CUADERNOS
DE LA FUNDACIÓN

C / 205

Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference

Montserrat Guillén
Ángel A. Juan
Helena Ramalhinho
Isabel Serra
Carles Serrat
(editors)

FUNDACIÓNMAPFRE

Área de Seguro y Previsión Social

Área de Seguro y Previsión Social

Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference

Montserrat Guillén
Ángel A. Juan
Helena Ramalhinho
Isabel Serra
Carles Serrat
(editors)

FUNDACIÓNMAPFRE

FUNDACIÓN MAPFRE no se hace responsable
del contenido de esta obra, ni el hecho de publicarla implica
conformidad o identificación con la opinión
del autor o autores.

Prohibida la reproducción total o parcial de esta obra
sin el permiso escrito del autor o del editor.

© 2015, FUNDACIÓN MAPFRE
Paseo de Recoletos, 23
28004 Madrid (España)

www.fundacionmapfre.org
publicaciones.ics@fundacionmapfre.org

ISBN: 978-84-9844-496-4
Depósito Legal: M-17626-2015
Maquetación y producción editorial: Cyan, Proyectos Editoriales, S.A.

PRESENTATION

Since 1975 FUNDACIÓN MAPFRE has been carrying out general interest activities for society in different professional and cultural areas, in addition to many actions aimed at improving the economic and social conditions of the most disadvantaged people and sectors of society.

We work within the Insurance and Social Protection Area with the aim of promoting and disseminating knowledge on, and the cultures of, Insurance and Social Protection.

In terms of the activities aimed at society in general, we create free and universal content on insurance, which we make available via the *Seguros y Pensiones para Todos (Insurance and Pensions for Everyone)* website. We organize training courses and educational and awareness-raising activities for teachers, as well as workshops for schoolchildren and free group visits to the Insurance Museum. We also publish informative guides to help people learn the basics of insurance.

In addition to this informative work, we support research by preparing reports on insurance markets and other subjects of interest, awarding research grants related to insurance and social protection, publishing books and surveys on insurance-related matters and organizing conferences and seminars. Our commitment to knowledge is embodied in our specialized Documentation Center that supports all our activities and is open to the general public.

These activities include the publication of this book, which contains articles that were presented at the *International Conference on Risk Analysis ICRA 6 / RISK 2015*, held in Barcelona, from May 26 to 29, 2015.

All information regarding our activities is available online and accessible to users around the world quickly and efficiently via our website: www.fundacionmapfre.org

Insurance and Social Protection Area. FUNDACIÓN MAPFRE

PRESENTACIÓN

Desde 1975 FUNDACIÓN MAPFRE desarrolla actividades de interés general para la sociedad en distintos ámbitos profesionales y culturales, así como acciones destinadas a la mejora de las condiciones económicas y sociales de las personas y de los sectores menos favorecidos de la sociedad.

Desde el Área de Seguro y Previsión Social trabajamos con el objetivo de promover y difundir el conocimiento y la cultura del Seguro y la Previsión Social.

En cuanto a las actividades orientadas hacia la sociedad en general, creamos contenidos gratuitos y universales en materia de seguros que divulgamos a través de la página web *Seguros y Pensiones para Todos.* Organizamos actividades educativas y de sensibilización mediante cursos de formación para el profesorado, talleres para escolares y visitas gratuitas para grupos al Museo del Seguro. Asimismo publicamos guías divulgativas para dar a conocer aspectos básicos del seguro.

Además de esta labor divulgativa, apoyamos la investigación mediante la elaboración de informes sobre mercados aseguradores y otros temas de interés, la concesión de ayudas para la investigación en seguros y previsión social, la publicación de libros y cuadernos de temática aseguradora y la organización de jornadas y seminarios. Nuestro compromiso con el conocimiento se materializa en un Centro de Documentación especializado que da soporte a todas nuestras actividades y que está abierto al público en general.

Dentro de estas actividades se encuadra la publicación de este libro, que recoge los artículos presentados en la International Conference on Risk Analysis ICRA 6/RISK 2015, celebrada en Barcelona del 26 al 29 de mayo de 2015.

Todas nuestras actividades se encuentran disponibles y accesibles en internet, para usuarios de todo el mundo, de una manera rápida y eficaz, a través de nuestra página web: www.fundacionmapfre.org.

Área de Seguro y Previsión Social. FUNDACIÓN MAPFRE

EDITORS

Montserrat Guillén, Universitat de Barcelona, Spain
Ángel A. Juan, Open University of Catalonia, Spain
Helena Ramalhinho Lourenço, Universitat Pompeu Fabra, Spain
Isabel Serra, Universitat Autònoma de Barcelona, Spain
Carles Serrat, Technical University of Catalonia, Spain

GENERAL CHAIRS

Montserrat Guillén, Universitat de Barcelona, Spain
Ángel A. Juan, Open University of Catalonia, Spain
Christos Kitsos, Technological Educational Institution of Athens, Greece
Teresa Oliveira, Universidade Aberta, Portugal
Helena Ramalhinho Lourenço, Universitat Pompeu Fabra, Spain
Isabel Serra, Centre de Recerca Matemática, Spain
Carles Serrat, Technical University of Catalonia, Spain

INTRODUCTION

The 2015 International Conference on Risk Analysis (ICRA6) is an international forum for disseminating recent advances in the field of Risk Analysis, with applications for the risk assessment and the risk management in a variety of fields: Life, Biology, Environmental Sciences, Public Health, Economics, Insurance, Finance, Reliability of Engineering and Technical, Biological & Biomedical Systems. The conference also includes the 6th Workshop on Risk Management and Insurance (RISK2015) which has been taken place biannually in Spain.

ICRA6 and RISK2015 were held in Barcelona and the sessions took place at the Universitat Pompeu Fabra on May 26-29, 2015.

There were over 100 submissions. A selection of short papers is included in this book. All of them went through a double peer-review process.

The members of the organizing committee would like to thank the scientific committee for their valuable help to make this conference a success. We want to acknowledge participants, presenters, chairpersons and all authors in general for their contribution to the development of new advances in the analysis of risk. We also want to thank our invited speakers Prof. Dr. Sebastian Martorell (Universitat Politècnica de València, Spain), Prof. Dr. Mei-Ling Tin Lee (University of Maryland, USA) and Prof. Dr. Hansjörg Albrecher (University of Lausanne, Switzerland) for accepting to lay a central role in this conference.

We are indebted to Fundacion Mapfre for sponsoring this publication and to all other sponsors for their generous support.

Barcelona, May 2015

INTERNATIONAL SCIENTIFIC COMMITTEE

Pilar Abad Romero, Universidad Rey Juan Carlos, Spain
Mercedes Ayuso Gutiérrez, Universidad de Barcelona, Spain
Alejandro Balbás de la Corte, Universidad Carlos III de Madrid, Spain
David Banks, Duke University, USA
Inmaculada Barranco-Chamorro, Universidad de Sevilla, Spain
Catalina Bolancé, Universidad de Barcelona, Spain
Christian Brownlees, Universitat Pompeu Fabra, Spain
Alejandra Cabaña, Universitat Autonòma de Barcelona, Spain
Chrys Caroni, National Technical University of Athens, Greece
Jalila Daoudi, Unversitat Pompeu Fabra, Spain
Ana Debon, Universitat Politècnica de València, Spain
José M. Feria Domínguez, Universidad Pablo de Olavide, Spain
Jerzy Filus, Oakton Community College, USA Lidia Filus, Northeastern Illinois University, USA
Ilia Frenkel, Ben-Gurion University of the Negev, Israel
Anna Gordon, Mindjet Inc., USA
Luís Grilo, Instituto Politécnico de Tomar, Portugal
Sneh Gulati Gulati, Florida International University, USA
Emilio Gómez Déniz, Universidad de Las Palmas de Gran Canaria, Spain
George Halkos, University of Thessaly, Greece
Antonio J. Heras Martínez, Universidad Complutense de Madrid, Spain
Luisa Huaccho, University of Leeds, UK Catherine Huber, Universite Paris Descartes, France
Maria Ivette Gomes, Faculdade de Ciências da Universidade de Lisboa, Portugal
Enrique J. Jiménez Rodríguez, Universidad Pablo de Olavide, Spain
Alex Karagrigoriou, University of the Aegean, Cyprus Renatas Kizys, University of Portsmouth, UK
Andrei Korobeinikov, CRM, Spain
Víctor Leiva, Universidad Adolfo Ibáñez, Chile
Roberval Lima, Embrapa Brasil e Universidade Federal do Amazonas, Brasil
Nikolaos Limnios, University of Technology of Compiègne, France
Mei Ling Ting Lee, University of Maryland, USA Edwirde Luiz Silva, Universidade Estadual de Paraíba, Brasil

Maria Manuela Neves, Instituto Superior de Agronomia - Universidade Técnica de Lisboa, Portugal
Sebastian Martorell, Universitat Politècnica de València, Spain
David Moriña Soler, Centre for Research in Environmental Epidemiology, Spain
Tim Myers, CRM, Spain
Tsuyoshi Nakamura, Nagoya Institute of Technology, Japan
Vicente Núñez-Antón, Universidad del País Vasco, Spain
Amílcar Oliveira, Universidade Aberta, Portugal
Lluís Ortiz, CRM, Spain
Dinis Pestana, Faculdade de Ciências da Universidade de Lisboa, Portugal
Paul Rochet, Université de Nantes, France
Juan Manuel Rodríguez Poo, Universidad de Cantabria, Spain
José María Sarabia Alegría, Universidad de Cantabria, Spain
Bartosz Sawik, AGH University of Science and Technology, Poland
Tadeusz Sawik, AGH University of Science and Technology, Poland
Miguel Ángel Sordo Díaz, Universidad de Cádiz, Spain
Milan Stehlík, Johannes Kepler University, Austria
José Luis Vilar Zanón, Universidad Complutense de Madrid, Spain
Francisco José Vázquez Polo, Universidad de Las Palmas de Gran Canaria, Spain

TABLE OF CONTENTS

1. European Financial Integration in Times of Crisis
Pilar Abad and Helena Chulia 23

2. An Actuarial Estimation of the Individual Cost of Retirement and Widowhood Pension: An Application to the Spanish Case
Estefanía Alaminos and Mercedes Ayuso 33

3. On Competitive Non-Life Insurance Pricing under Incomplete Information
Hansjoerg Albrecher and Dalit Daily-Amir 41

4. El perfil del conductor en estado de alcoholemia: Diferencias entre carretera y zona urbana
Manuela Alcañiz, Miguel Santolino and Lluís Ramon 49

5. Pricing Life Insurance with Triangular Interest Rates and Fuzzy Random Variables: Some Analytical Results for Mixed Endowments
Jorge de Andrés-Sánchez and Laura González-Vila Puchades 57

6. Alternative Bounded Distributions to Model Chinese Financial Returns
Francisco Aranda, Catalina García and José María Pérez Sánchez 65

7. Comparison of Reserving and Spatial Models in Insurance
Miklós Arató and László Martinek 71

8. Modeling Bivariate Kumaraswamy Dependent Risks
Barry C. Arnold and Indranil Ghosh 79

9. On a New Multivariate IFR Notion for Random Vectors Based on the Standard Construction
Antonio Arriaza-Gómez, Félix Belzunce, Julio Mulero and Alfonso Suárez-Llorens 87

10. Gender Differences in the Effect of Driving Patterns on the Risk of Accident in PAYD Insurance
Mercedes Ayuso and Ana María Pérez-Marín 95

11. A Model for Rating Transitions over Time
Sergio Baena Mirabete and Pere Puig 103

12. Some Parametric Tests in Partially-Non-Regular Log-Exponential Models
Inmaculada Barranco-Chamorro 111

13. Risk in Eco-innovation Introduction at Enterprises
Pawel Bartoszczuk 119

14. Risk Behavior Behind Distortion Risk Measures
Jaume Belles-Sampera 127

15. Analyzing and Managing Political Risk
Ole Bergfjord 135

16. A Bivariate INAR(1) Model for Insurance Claim Counts
Lluís Bermúdez, Montserrat Guillén and Dimitris Karlis 143

17. A Finite Mixture of Multiple Poisson Distributions for Modelling Days of Absence from Work due to Motor Accident
Lluís Bermúdez, Dimitris Karlis and Miguel Santolino 153

18. Influencia del reaseguro y de la estructura del colectivo en la solevencia de una cartera de vida
Eudald Boira, M.ª Àngels Pons and F. Javier Sarrasí 161

19. Claim Reserving: Calendar Year Reserves for the GLM
Eva Boj and M.ª Teresa Costa 169

20. An Approach to Solve the Weakness of Block Maxima Method
Alejandra Cabaña and Isabel Serra 179

21. Solving the Multi-Depot Vehicle Routing Problem Considering Uncertainty and Risk Factors
Laura Calvet, Ángel A. Juan and Nina Schefers 187

22. First Hitting Time Regression for Recurrent Events with Random Effects
Chrys Caroni, Sonia Malefaki and Polychronis Economou 195

23. Non-stationary Modelling of Extremes of Precipitation and Temperature over Mountainous Areas under Climate Change
Chrys Caroni, Dionysia Panagoulia and Polychronis Economou 203

24. Modeling Trends and Inequality of Longevity in the European Union Countries
Patricia Carracedo, Ana Maria Debon and Steve Haberman 211

25. Optimal Designs Subject to Cost Constraints in Simultaneous Equations Models
Víctor Casero-Alonso and Jesús López-Fidalgo 219

26. A Numerical Analysis of Risk Measures under Proportional Reinsurance
Anna Castañer, M. Mercè Claramunt and Maite Mármol 229

27. Modeling Longevity Risk with Generalized Dynamic Factor Models and Vine-Copulas
Helena Chuliá, Montserrat Guillén and Jorge Uribe 237

28. Performance of the Clauset-Shalizi-Newman Method for Power-law Fitting and Comparison with other Methods
Álvaro Corral and Isabel Serra 245

29. DPLN: Data Analysis in Finance and Insurance.
Jalila Daoudi and Rosa Lillo 251

30. Modeling the IBEX-35 Index Using a Compound Poisson Stochastic Process
Ana Debón, Juan Carlos Cortés, Óscar Monzó-Chafer and Fabián Rodríguez-Santos 257

31. Measuring the Systemic Risk by Multivariate Indicators
Elena Di Bernardino, José María Fernández-Ponce, Fátima Palacios Rodríguez and María del Rosario Rodríguez-Griñolo 265

32. Estimation of Multivariate Critical Layers: Applications to Rainfall data
Elena Di Bernardino and Didier Rullière 273

33. Financial Time Series obtained by Agent-based Simulation
Pol Ferrando and Josep Fortiana 281

34. Euler-Poisson Schemes for Lèvy Processes
Albert Ferreiro-Castilla 289

35. On some Generalizations of the Univariate Normal Density
Jerzy Filus 297

36. Construction of Conditional Probability Densities as Enforced Regression
Jerzy Filus and Lidia Filus 303

37. Risk of Data Loss on QR Codes: Hadamard Matrices and Links to Block Designs
Carla Francisco and Teresa Oliveira 311

38. Components with Higher and Lower Risk in a Reliability System
Josep Freixas and Montserrat Pons 319

39. Uncertainty Analysis in the IS-LM Model
Joaquim Gabarro and Maria Serna 329

40. Dynamic Shape Analysis - before and after - Chemotherapy
Giebel Stefan, Philipp Hermann, Milan Stehlik and Jens-Peter Schenk 339

41. Predicting wind Speeds for Energy Risk: Weibull Distribution and Neural Network Methods as Alternative and Complementary Strategies
Giebel Stefan and Martin Rainer 347

42. Extensions and Applications of CompARE: web Platform to select the Primary Endpoint in a Randomized Clinical Trial
Moisés Gómez Mateu and Guadalupe Gómez Melis 357

43. Selection of the Primary Endpoint in a Clinical Trial: Composite Binary Endpoints versus Single Outcomes
Guadalupe Gómez Melis and Aintzane Ayestaran 367

44. The Exponential-Reciprocal Generalized Inverse Gaussian Distribution. A New Model to Describe Loss Distributions
Emilio Gómez-Déniz, Francisco-José Vázquez-Polo, Victoriano García and Enrique Calderín-Ojeda 377

45. Determination of the Optimal Replacement Age for a Preventive Maintenance Problem involving a Weibull Failure Probability Distribution Function
Ernesto Gutiérrez-Miravete 385

46. On the Stability of the Reforestation Optimal Control Model Assumption
David Hampel, Jitka Janova and Jiri Kadlec 391

47. Oscillatory Modes Organization in Complex Temporal Behaviors
Ramón Herrero, Gaspar Orriols and Francesc Pi 399

48. Fighting the Ebola Virus: An Exercise of Risk Analysis and Optimal Decision Making
John Houston and Madhu Acharyya 407

49. Survival and Quality of Life
Catherine Huber 415

50. Poverty as a Risk Factor for Sustainable Development
Ladislava Issever Grochová 423

51. Measuring Risk in Costing Systems Using Costing at Risk (CaR)
Víctor Jiménez and Paulo Afonso 431

52. On Risk Process with Pareto Mixing
Pavlina Jordanova, Yulia Nefedova and Milan Stehlík 439

53. On the Michaelis–Menten Model
Christos Kitsos and Thomas Toulias 449

54. Solving the Portfolio Optimization Problem under Realistic Constraints
Renatas Kizys, Angel Juan, Bartosz Sawik and Ángel Linares 457

55. Effective Risk Assessment. Review - Power Utilities
Pui Yong Leo 465

56. Silver Supply Risk Assessment For Sustainable Photovoltaic Installation
Wenhua Li and Tsuyoshi Adachi 473

57. Nueva propuesta de ordenación de compañías aseguradoras
Salvador Linares-Mustarós, José María Merigó-Lindahl and Anna María Gil-Lafuente 483

58. Storm Surge Hazard Estimation on the United States' Atlantic Coast Using the Clustered Separated Peaks-Over-Threshold Simulation (CSPS) Method
Madeleine Lopeman, George Deodatis and Guillermo Franco 491

59. Multivariate Density and Derivative Gamma Kernel Estimation by Dependent Data
Liubov Markovich 499

60. Hitting Times of Threshold Exceedances and Their Distributions
Natalia Markovich 507

61. Interpreting Textual User Generated Content for Supplier-Credit Risk Analysis
Aaron Mengelkamp, Sebastian Hobert and Matthias Schumann 515

62. Estimating Cumulated Absorbed Doses and Associated Health Risks due to Occupational Exposure to Ionizing Radiation
David Moriña Soler, James Grellier, Adela Carnicer, Eileen Pernot and Elisabeth Cardis 525

63. Extreme Value Analysis of River Levels in a Hydrometric Station in the North of Portugal
Manuela Neves, Helena Penalva and Sandra Nunes 533

64. On Measures of Financial Risk
S. Novak 541

65. Stochastic Response Surface Methodology: A Tool to Risk Assessment
Teresa Oliveira, Conceição Leal and Amílcar Oliveira 551

66. An Approach for Risk Estimation in Territories/Perils Un-modelled by the Insurance Industry
Delioma Oramas Dorta 561

67. Alternative Methods of Estimating the Longevity Risk
Arelly Ornelas and Catalina Bolancé 571

68. Extreme Value Theory in Computer Sciences: The Case of Embedded Safety-critical Systems
María Padilla, Jaume Abella, Joan del Castillo and Francisco J. Cazorla 579

69. Considering Risk Issues in the Uncapacitated Facility Location Problem with Stochastic Components
Adela Pagès Bernaus, Ángel A. Juan and Helena Ramalhinho 587

70. On Several Generalizations of Classical Tail Index Estimators
Vygantas Paulauskas and Marijus Vaičiulis 597

71. Flexible Bonus-Malus Systems Induced by OWA Optimization
José María Pérez Sánchez and Víctor Blanco Izquierdo 605

72. The Effect of Zero-inflation in Claims Insurance Through a Bayesian Model
José María Pérez Sánchez and Emilio Gómez Déniz 613

73. Forecasting Probability of Bankruptcy from Unbalanced Data
Francesca Pierri, Elena Stanghellini and Nicolò Bistoni 621

74. Joint Modeling Scheme by Weighting Cumulative Effects over Time. Research on Mortality for Insured Elderly
Xavier Piulachs and Ramon Alemany 629

75. Real Options Valuation under Uncertainty
Federico Platania, Marie Lambert and Manuel Moreno 637

76. Behavioral Equilibrium Exchange Rates: Re-estimation for the Euro Area Countries
Klára Plecitá and Ladislava Issever Grochová 647

77. Solving the Integrated Location Routing Problem Considering Uncertainty and Risk factors
Carlos Quintero-Araujo and Ángel A. Juan 655

78. Capital Allocation under Dependent Risks: An Application to GHG Emissions
Lorena Remuzgo, José María Sarabia and Carmen Trueba 663

79. A Test on Separate Families of Hypotheses: Distinguish Between Uniform and Exponential Distributions
Paul Rochet and Isabel Serra 671

80. Modeling *N* Electrical Component Failures using Non-stationary Markovian Arrival Processes
Joanna Rodriguez, Rosa Lillo and Pepa Ramírez-Cobo 677

81. La distribución Pareto Positive Stable, un modelo estocástico para eventos extremos
Antonio José Sáez-Castillo, José María Sarabia and Faustino Prieto 687

82. Risk Aggregation in a Class of Dependent Heavy-tailed Distributions
José María Sarabia, Faustino Prieto and Vanesa Jordá 695

83. Statistical Tests for the Tail of the Seismic-moment Distribution of Global Shallow Earthquakes
Isabel Serra and Álvaro Corral 703

84. A Shared-Parameter Joint Model for Prostate Cancer Risk and PSA Longitudinal Profiles
Carles Serrat, Montserrat Rué, Carmen Armero, Xavier Piulachs, Hèctor Perpiñán, Anabel Forte, Álvaro Páez and Guadalupe Gómez 711

85. The Theory of Economic Systems as base for Structuring Space of Risk Factors
Yulia Sleptsova and Roman Kachalov 721

86. Cost Risk Analysis Methods for Defence Acquisition Projects
Abderrahmane Sokri and Ahmed Ghanmi 729

87. On multivariate Extensions of the Increasing Convex Order to Compare Risks
Miguel A. Sordo 737

88. Fractal Based Cancer Modelling
Milan Stehlík 741

89. Notes on Robust Testing for Normality in Insurance Science
Luboš Střelec 749

90. An Introduction to the Safety Evaluation of Ingredients in Household Chemical Products
Jin-Hyang Suh, Myung-Sil Hwang and In Gyun Hwang 757

91. Childhood Obesity as Cardiovascular Disease Risk Predictor
Lucía Tarro, Elisabet Llauradó Ribé, Magaly Aceves Martins, David Moriña, Montse Giralt Batista and Rosa Solà Alberich 765

92. Validation of Risk Scoring Models for Trade Credit in Small and Medium Enterprises
Manuel Terradez, Ana Debon and Ángel A. Juan 773

93. Risk Analysis Using Threshold Regression
Mei-Ling Ting Lee 781

94. Stochastic Bounds for Times Elapsed between Successive Failures of Components in Reliability Systems
Nuria Torrado and Rosa E. Lillo 787

95. Multivariate Risk Measures: A Directional Approach for Value at Risk
Raúl Andrés Torres Díaz, Rosa E. Lillo and Henry Laniado 795

96. Local Maximum Based Tail Index Estimator
Marijus Vaičiulis 803

97. Application of Stochastic Modeling to Online Product Returns Risk Analysis
José L. Vilar, Eduardo Vilar and Antonio Heras 811

98. Decomposition of Hair Mineral Variations into Intra-individual, Inter-individual and Physical Variations
Tomomi Yamada, Koshi Kataoka, Todd Saunders, Koichiro Sera, Toshihiro Takatsuji, Tsuyoshi Nakamura and Yoshiaki Nose 817

99. On a Shape-Invariant Hazard Regression Model
Cheng Zheng and Ying Qing Chen 827

List of authors 835

1. EUROPEAN FINANCIAL INTEGRATION IN TIMES OF CRISIS

P. Abad
University Rey Juan Carlos and Riskcenter-IREA,
Departament of Economic, Spain
pilar.abad@urjc.es

H. Chulià
University of Barcelona and Riskcenter-IREA,
Department of Econometrics, Spain
hchulia@ub.edu

ABSTRACT

In this paper we investigate the dynamics of European government bond market integration during the financial crisis and, subsequently, during the European sovereign debt crisis. Based on the approach developed by Bae *et al.* (2003), we adopt an intuitive measure of integration: the higher the number of joint extreme price rises or falls (coexceedances), the higher the degree of integration. Our results reveal that the level of integration of European government bond markets with the euro area has changed over time, with notable differences between the financial and the European sovereign debt crises.

1. INTRODUCTION

The extent to which European government bond markets are integrated is a key question for policymakers and market participants. Policymakers are particularly keen to understand the mechanisms that link these markets in order to be able to make effective monetary policy decisions and to maintain financial stability. Likewise, an understanding of bond market linkages can help market participants formulate appropriate

risk management strategies and investment decisions. This interest becomes even greater in years of turmoil when financial markets are hit by extreme shocks.

In the financial literature, a wide variety of frameworks have been employed for the empirical examination of the integration of European government bond markets. The central focus of early papers was on the role that the European Monetary Union (EMU) played in the process of financial integration of the EU-15 bond markets. In this line of research, some studies have assessed the relative importance of systemic and idiosyncratic risk in EMU sovereign yield spreads (see Geyer *et al.*,2004; Pagano and von Thadden, 2004; Gomez-Puig, 2009a and 2009b). A different perspective is provided by Christiansen (2007) who show that the bond markets of EMU countries are more integrated than those of non-EMU countries and that these markets became more integrated following the introduction of the euro. A more recent study by Beber *et al.* (2009) finds that the bulk of sovereign yield spreads can be explained by differences in credit quality, though liquidity plays a non-trivial role. Finally, Abad *et al.* (2010) find that euro area countries are only partially integrated and present differences in their market liquidity and default risk. They also find that the markets of countries sharing a monetary policy are more vulnerable to regional risk factors and that the countries that opted to stay out of the Monetary Union are more vulnerable to global risk factors. Another line of empirical research of European government bond market integration incorporates the new EU members into the analysis. Kim *et al.* (2005) find that the degree of integration of the new members with the German bond market is weak and stable, with little evidence of any further strengthening despite increased political integration. Cappiello *et al.* (2006) only document an increase in integration for the Czech Republic's bond market versus Germany's.

Finally, a new line of research investigates the impact of the financial crisis on European government bond market integration. Von Hagen *et al.* (2011) find that the larger spreads observed during the financial crisis are the result of a higher penalty imposed by the markets on fiscal imbalances and of greater international risk aversion. The results of Pozzi and Wolswijk (2012) suggest that the idiosyncratic factors were almost eliminated in all countries by 2006 but subsequently reappeared as a consequence of the financial crisis. Abad *et al.* (2014) show that, from the onset of the financial market tensions in August 2007, markets moved towards higher segmentation, and the differentiation of country risk factors increased substantially

across countries. Christiansen (2014) concludes that the integration of EMU members has not been so great during the recent crisis.

This study assesses the integration of a selected number of European government bond markets with the euro area during periods of turbulence, when investors and policymakers have a particularly strong interest in whether and how shocks propagate to other countries. Following the launch of the euro in January 1999, the markets priced the debt of the European member states as being virtually identical. In the period 2003-2007, spreads remained very small and did not reflect differences in the fiscal positions of the countries. As such, the period was characterised by a significant underpricing of risk, leaving investors to search for yield in an environment of abundant global liquidity. This progress towards financial integration was interrupted and reversed, however, by the global financial crisis and, more recently, by the European sovereign debt crisis, in which sovereign bond markets have been dominated by sharp differentiation, especially across borders.

Based on the approach developed by Bae *et al*. (2003), we measure the integration of each European government's bond market with the euro area by examining how often extreme returns on each bond market and the euro area occur simultaneously. This analysis provides helpful information on the dynamics leading to joint extreme price rises or falls and allows us to adopt an intuitive measure of integration: the higher the number of coexceedances with the euro area, the higher the degree of integration. Bae *et al*. (2003) capture the coincidence of extreme return shocks across countries within a broader region and also across regions. They define contagion within regions as the fraction of coexceedances that cannot be explained by fundamentals and contagion across regions as the fraction of coexceedances unexplained by fundamentals that is explained by the exceedances from other countries. In this paper, we are interested in analysing the integration of a selected number of European government bond markets with the euro area and in testing whether there are differences across countries with respect to the underlying determinants of the level of integration. To this end, we carry out a hierarchical cluster analysis that allows us to group countries in terms of their level of integration over the sample period.

We address two basic sets of questions. First, how closely are the European government bond markets associated with the euro area? And, has the level of

integration of these markets changed during the recent years of turmoil? It is intuitive that financial market integration changes with economic conditions. The main results of this paper can be summarized as follows. We find that the level of integration of European government bond markets with the euro area has changed over time and that the bond markets analysed group differently over the sample period in terms of integration.

2. EMPIRICAL ANALYSIS

The data consist of the ten-year JPMorgan Government Global Bond Index (JPMGBI), expressed in terms of a common currency, the euro, and the sample includes 16 European countries. Our study focuses on ten EMU EU-15 countries (Austria, Belgium, France, Germany, Greece, Ireland, Italy, the Netherlands, Portugal and Spain) and six non-EMU countries (Denmark, the Czech Republic, Hungary, Poland, Sweden and the UK). As a proxy for the entire euro area we use the JP Morgan EMU Government Index. These bond market indices are transformed into returns by taking the first difference of the natural log of each bond price index. All data have been collected from Thomson Datastream.

We use daily data for the period January 2005 through December 2013, thus our sample covers the recent years of turmoil (initially the financial crisis and, subsequently, the European sovereign debt crisis). We define the starting point of the financial crisis as August 2007, when equity markets initially fell and central banks started intervening to provide liquidity to financial markets. For our analysis, we match the end of the financial crisis with the beginning of the European sovereign debt crisis. As pointed out by Christiansen (2014), dating the European sovereign debt crisis is not a straightforward task as no official dates are available. Generally, it is considered to have begun in late 2009 and is still running its course. Therefore, we define the starting point of the sovereign crisis as January 2010 and it runs till the end of our dataset in December 2013.

Following Bae *et al.* (2003), we define an extreme return, or exceedance, as one that lies either below (above) the 5th (95th) quantile of the marginal return distribution. Similarly, we define a coexceedance as the occurrence of extreme

returns in one European government bond market and in the euro area simultaneously on a given day; thus, the higher the number of coexceedances with the euro area, the higher the degree of integration. Given the diversity of economic and financial structures across the EU economies, the fact that not all the countries belong to the EMU and that some countries only became members of the EU relatively recently, it is standard in the literature to divide European countries into four groups: (1) EMU EU-15 central countries (Austria, Belgium, France, Germany and the Netherlands), (2) EMU EU-15 peripheral countries (Greece, Ireland, Italy, Portugal and Spain), (3) Non-EMU new EU countries (the Czech Republic, Hungary and Poland), and (4) non-EMU EU-15 countries (Denmark, Sweden and the UK).

However, the recent years of turmoil might have produced heterogeneity within groups or even homogeneity between countries in different groups in terms of their respective levels of integration. To analyse this possibility, we carry out a hierarchical cluster analysis that enables us to group countries that present similar characteristics across a set of variables. Here, this set of variables refers exclusively to the coexceedances of each government bond market with the euro area over time, i.e. its level of integration. The end result is a map (dendrogram) that allows us to visualize the groups. In so doing, we are able to test whether the cluster analysis (in terms of the level of integration) leads to the same classification of countries as described above. As mentioned, from the introduction of the common currency until the end of 2007, the bonds of the EMU countries were almost perfect substitutes but this situation changed, first, with the financial crisis and, subsequently, with the European sovereign debt crisis. For this reason, to determine whether the cluster groups have been stable over the sample period, we perform the cluster analysis for three sub-samples: the "tranquil period", the financial crisis and the European sovereign crisis.

Figure 1a shows that during the "tranquil period", the central countries (Austria, Belgium, France, Germany and the Netherlands) form a cluster, to which Italy is added. The dendrogram also shows the similarity between the peripheral bond markets (Greece, Ireland, Portugal and Spain), with the exception of Italy, and to which Denmark, Hungary and Poland are added. Finally, the UK, Sweden and the Czech Republic form a separate alignment.

The picture changes somewhat when we consider the financial crisis (Figure 1b). As in the "tranquil period", the central countries form a cluster, to which Spain and Denmark are now added. The Czech Republic and Poland, the new EU members, cluster together. The similarities between Greece, Ireland and Portugal, the first peripheral economies to collapse, can be clearly identified and they form a group together with Italy. The UK and Sweden continue to be independent of the other countries, as now is Hungary.

Finally, Figure 1c shows that during the European sovereign debt crisis, as expected, the central countries once more form a group (together with Denmark). Italy and Spain cluster together, as do the new members of the EU together with Greece, Ireland and Portugal. This result indicates that the peripheral bond markets are divided into two groups: those most affected by the European sovereign debt crisis (Greece, Ireland and Portugal), whose levels of integration fall to levels similar to those of the new member states; and Italy and Spain, which remain at higher levels of integration. Sweden and the UK also form a separate alignment.

Overall, the results of the clustering analysis suggest that the level of integration of European government bond markets with the euro area, measured in terms of coexceedances, has changed over the three sub-samples under analysis. Some government bond markets present a certain degree of instability indicating that the effects of the crises have not been homogeneous across these countries. As a consequence, opportunities for diversification have changed over the sample period.

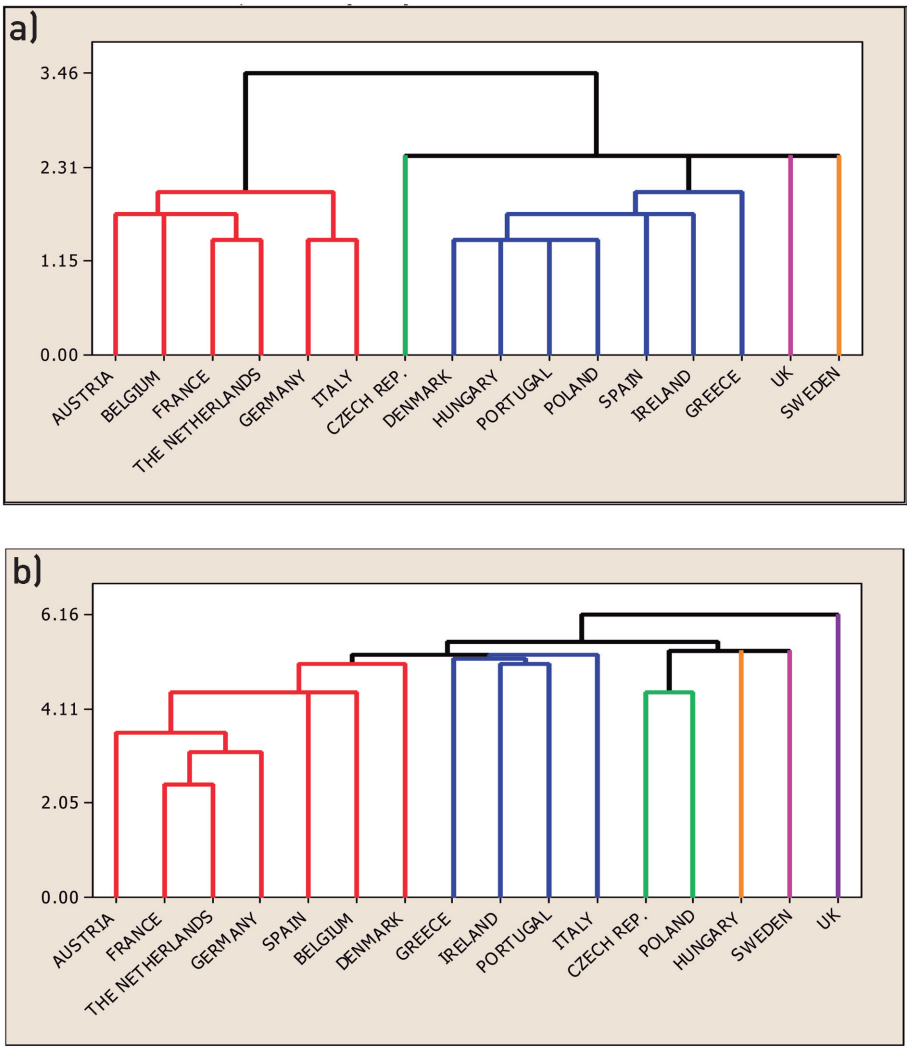
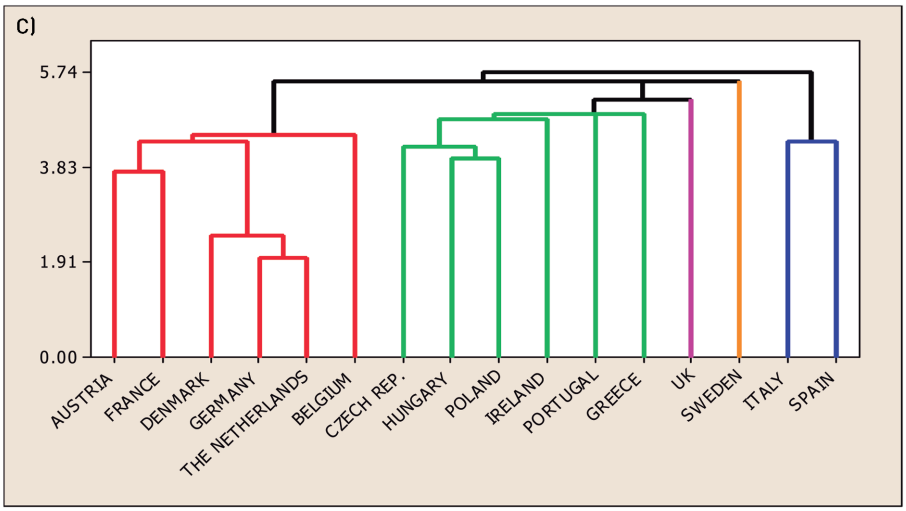


Figure 1. Hierarchical Cluster Analysis

a) "Tranquil" period: 1/1/2005 to 8/6/2007

b) Financial crisis: 8/7/2007 to 12/31/ 2009

c) European sovereign debt crisis: 1/1/2010 to 12/15/2013



Note. The dendrogram is based on the single amalgamation method with a Euclidean distance measure.

3. CONCLUSION

We report evidence that the degree of integration of European government bond markets has changed over the sample period: first, during the financial crisis and, subsequently, during the European sovereign debt crisis. Moreover, the effects of these crises have not been the same across all the bond markets, to the extent that the traditional groupings of markets on the basis of the level of integration vary across the three subsamples.

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Ministry of Economy and Competitiveness (ECO2011-23959 and ECO2012-35584).

REFERENCES

Abad, P., Chuliá, H., and Gómez-Puig, M. (2014). "Time-varying integration in European Government Bond Markets". *European Financial Management*, 20(2), 270-290.

Abad, P., Chuliá, H., and Gómez-Puig, M. (2010). "EMU and European Government Bond Market Integration". *Journal of Banking and Finance*, 34, 2851-2860.

Bae, K.-H., Karolyi, G.A., and Stulz, R.M. (2003). "A New Approach To Measuring Financial Contagion". *Review of Financial Studies*, 16 (3), 717-763.

Beber, A., Brandt, M.W., and Kavajecz, K.A. (2009). "Flight-to-Quality or Flight-to-Liquidity? Evidence from the Euro-Area Bond Market". *Review of Financial Studies,* 22(3), 925-957.

Cappiello, L., Gerard, B., Kadarenja, A., and Mangenelli, S. (2006). *Financial Integration of New EU Member States,* Working Paper, ECB.

Christiansen, C. (2007). "Volatility-spillover Effects in European Bond Markets". *European Financial Management,* 13(5), 923-948.

Christiansen, C. (2014). "Integration of European Bond Markets". *Journal of Banking and Finance*, 42, 191-198.

Geyer, A., Kossmeier, S., and Pichler, S. (2004). "Measuring Systematic Risk in EMU Government Yield Spreads". *Review of Finance,* 8, 171-197.

Gómez-Puig, M. (2009a). "The Immediate Effect Of Monetary Union Over Eu-15'S Sovereign Debt Yield Spreads". *Applied Economics,* 41, 929-939.

Gómez-Puig, M. (2009b). "Systemic and Idiosyncratic Risk in EU-15 Sovereign Yield Spreads after Seven Years of Monetary Union". *European Financial Management,* 15, 971-1000.

Kim, S. J., Lucey, B., and Wu, E. (2006). "Dynamics of Bond Market Integration between Established and Accession European Union Countries". *Journal of International Financial Markets, Institutions and Money,* 16(1), 41-56.

Pagano, M., and von Thadden, E.L. (2004). "The European Bond Markets under EMU". O*xford Review of Economic Policy*, 20, 531-554.

Pozzi, L., and Wolswijk, G. (2012). "The Time-Varying Integration of Euro Area Government Bond Markets". *European Economic Review*, 56(1), 36-53.

Von Hagen, J., Schuknecht, L., and Wolswijk, G. (2011). "Government Bond Risk Premiums in the EU Revisited: The Impact of the Financial Crisis". *European Journal of Political Economy,* 27(1), 36-43.

2. AN ACTUARIAL ESTIMATION OF THE INDIVIDUAL COST OF RETIREMENT AND WIDOWHOOD PENSION: AN APPLICATION TO THE SPANISH CASE

Estefanía Alaminos
Universitat de Barcelona, Departament d'Econometria, Estadística i Economia Espanyola, Riskcenter, Spain, estefaniaalaminos@outlook.com

Mercedes Ayuso
Universitat de Barcelona, Departament d'Econometria, Estadística i Economia Espanyola, Riskcenter, Spain, mayuso@ub.edu

ABSTRACT

This paper analyses a Multiple State Actuarial Model for Married Widower with the aim to assess the expenditure that a retirement pensioner would cost to the Pension Scheme System in both cases, surviving as retired or surviving as retired and widower being entitled with the related pensions. We estimate the transition survival probability between marital states and the pure survival probabilities required for obtaining the expected present value in pensions by person.

1. INTRODUCTION

This paper examines two benefits provided by the Spanish Social Security System as part of its protection program: the contributory retirement pension and the widowhood pension. The retirement pension is a lifelong economic benefit received by beneficiaries who satisfy the requisite conditions, while the widowhood pension is a death/survivor benefit, its goal being to alleviate the necessity faced by an individual following the death of someone on whom they were financially dependent. As such

the person facilitating the right to the pension and the recipient are not the same. The widowhood pension is paid, conditional to the fulfillment of several requirements, when an individual affiliated to the Social Security System dies. Thus, the individual must have been an active or assimilated contributor, or the recipient of a retirement pension or a permanent disability pension. The recipient of the widowhood pension is the surviving spouse (or legally separated or divorced partner).

Recent studies question the present-day role of the widowhood pension (Hernández *et al.*, 2011; Kahale, 2011). The pension was introduced to alleviate the precarious situation faced by a widow on the death of her spouse; however, women now participate actively in the labor market and so are entitled to receive their corresponding contributory retirement pension.

This, linked with the fact that women enjoy an increasing life expectancy, means that in the future more and more women will receive both a retirement and a widowhood pension.

The most recent amendments to the pension law with regard to widowhood benefits have tightened access conditions for de facto spouses and divorcees, and raised the benefits for those on low incomes (Salvador, 2013). In general, across the EU member states, access conditions to pension benefits have been made stricter (Monticone *et al.*, 2008; CECS, 2010). Some studies have proposed the need for modifications to the Spanish widowhood pension, including the introduction of a model similar to that employed in Germany based on a pension-splitting arrangement (Hernández *et al.*, 2011; Kahale, 2011), allowing access to the benefit only to the person who was economically dependent on the spouse (Ahn *et al.*, 2007), or deducting the consumed benefits of the person who gave the right to the widowhood pension (Sanabria *et al.*, 2010).

As for the concurrent receipt of pensions, it is possible to receive a widowhood pension in conjunction with another benefit payment, the only restriction being a legal ceiling on the overall amount a beneficiary can receive (CECS, 2010).

The goal of the present article is to assess the individual costs of retirement and widowhood pensions. We estimate a life table according to marital status and use

the results to calculate the expected present value of total payments for both concepts. We assume that the individual is married at the age of retirement and so it is possible they will receive both pensions in the future.

2. A MULTIPLE STATE MODEL FOR THE RETIREMENT-WIDOWHOOD PENSION CASE

We use the same notation as employed by Haberman and Pitacco (1999) but we apply it to our married-widowhood case.

Being p_y^{cc} the annual probability of being married at age y + 1, p_y^{cv} the annual y y transition probability from the married to the widow/er status, and pvv the annual probability of being widow/er at age y+1. We denote by qc the annual probability of dying for a married individual aged y ($q_y^c = q_y^{cc} + q_y^{cv}$) and q_y^c the annual probability of dying for a widow/er person. The transition probabilities matrix is shown in Table 1 (upper triangular matrix).

	c	v	d
c	p_y^{cc}	p_y^{cv}	q_y^c
v	0	p_y^{vv}	q_y^v
d	0	0	1

Table 1. Transition matrix.

For $h = 1, 2, ...$, we have (Chapman-Kolmogorov relations):

$$ {}_hp_x^{cc} = {}_{h-1}p_u^{cc}p_{y+h-1}^{cc} $$

$$ {}_hp_x^{vv} = {}_{h-1}p_y^{vv}p_{y+h-1}^{vv} $$

$$ {}_hp_x^{cv} = {}_{h-1}p_y^{cv}p_{y+h-1}^{vv} + {}_{h-1}p_y^{cc}p_{y+h-1}^{cv} = \sum_{r=1}^{h}({}_{h-r}p_y^{cc}p_{y+h-r}^{cv}\prod_{g=1}^{r-1}p_{y+h-r+g}^{vv}) $$

where $r = 1, 2, \cdots, h$ y $g = 1, 2, \cdots, r-1$. According to this last expression, the probability of a married individual remaining married within a period of length h, ${}_hp_y^{cc}$,

can be split into the product of probabilities of a period less than h; the same applies to the survival of a widow/er, ${}_hp_y^{vv}$. Thus, the transition probability must consider all the possible states that the individual might experience.

Finally, we can obtain the expected present value of the total pension that a pensioner could receive either for the retirement pension alone, or for both the retirement and widowhood pensions, according to expression (1),

$$C_j(0,\omega) = \sum_{h=1}^{\omega}[aq^{h-1}{}_hp_y^{cc}v^h + (a+b)q^{h-1}{}_hp_y^{cv}v^h] \qquad (1)$$
$$= \sum_{h=1}^{\omega}[aq^{h-1}{}_hp_y^{cc}v^h + (a+b)q^{h-1}\sum_{r=1}^{h}({}_{h-r}p_y^{cc}p_{y+h-r}^{cv}\prod_{g=1}^{r-1}p_{y+h-r+g}^{vv})v^h]$$

where v^h is the discount factor ($v = (1+i)^{-1}$), a is the pension at retirement, b is the pension at widohood. Note that a and b are life annuities, growing in geometric progression with rate q (which is set by government annually[1]).

3. DATA AND PROBABILITY ESTIMATION

In this section we present the results obtained when applying the methodology outlined in the previous section to Spain's married and widow/er population. We draw on data from the Instituto Nacional de Estadística (INE) and the Ministerio de Empleo y Seguridad Social. Population data are obtained from the 2011 Population and Housing Census (INE). We select the married and widow population series aged between 65 and 100. We apply a smoothing spline[2] to the married series with the aim of smoothing a set of noisy observations[3].

[1] In Spain, this rate is determined by the law of Presupuestos Generales del Estado.

[2] For more details, see "Nonparametrics regression and generalized linear models: A roughness penalty approach" by Green and Silverman, 1993.

[3] Due to the Spanish Civil War (1936-1939) and the postwar period (1939-1959), Spain's demographics present certain fluctuations in the number of people born during these periods.

Raw death probabilities have been calculated by marital status using Death Statistics and Census population. Finally, the probabilities are adjusted using an exponential function.

The estimated life tables by marital status are included as Tables 2, 3 and 4, by total, male and female populations, respectively[4].

Married			Widow/er		Transition		
x	q_x^{cc}	p_x^{cc}	$q_x^{vv} = q_x^v$	$p_x^{vv} = p_x^v$	q_x^{cv}	p_x^{cv}	v_x
65-69	0.00761	0.95320	0.00700	0.99300	0.00014	0.03892	0.03905
70-74	0.01348	0.94837	0.01327	0.98673	0.00025	0.03764	0.03790
75-79	0.02388	0.96941	0.02513	0.97487	0.00008	0.00655	0.00663
80-84	0.04230	0.88499	0.04760	0.95241	0.00173	0.06925	0.07098
85-89	0.07495	0.80210	0.09015	0.90985	0.00554	0.11187	0.11742
90-95	0.13277	0.70765	0.17076	0.82924	0.01362	0.13233	0.14595
95-100	0.23522	0.64224	0.32345	0.67655	0.01982	0.08291	0.10273

Table 2. Life Table, total population 2011.
Source: Authors' compilation.

Married			Widow/er		Transition		
x	q_x^{cc}	p_x^{cc}	$q_x^{vv} = q_x^v$	$p_x^{vv} = p_x^v$	q_x^{cv}	p_x^{cv}	v_x
65-69	0.01063	0.95977	0.01969	0.98031	0.00029	0.02902	0.02931
70-74	0.01813	0.95843	0.03231	0.96769	0.00038	0.02269	0.02306
75-79	0.00038	0.96579	0.05302	0.94698	0.00009	0.00311	0.00320
80-84	0.05276	0.90654	0.08701	0.91299	0.00177	0.03716	0.03893
85-89	0.09001	0.82750	0.14279	0.85721	0.00589	0.07072	0.07661
90-95	0.15355	0.72406	0.23432	0.76568	0.01434	0.09371	0.10805
95-100	0.26195	0.40898	0.38453	0.61547	0.06327	0.20253	0.26580

Table 3. Life Table, male population 2011.
Source: Authors' compilation.

[4] Probabilities have been calculated for all studied ages. The tables displayed here only include the results for the lowest age in each of the five-year intervals.

Married			Widow/er		Transition		
x	q_x^{cc}	p_x^{cc}	$q_x^{vv} = q_x^v$	$p_x^{vv} = p_x^v$	q_x^{cv}	p_x^{cv}	v_x
65-69	0.00421	0.94797	0.00477	0.99523	0.00011	0.04760	0.04771
70-74	0.00792	0.94483	0.00958	0.99042	0.00023	0.04680	0.04703
75-79	0.01490	0.95532	0.01925	0.98075	0.00029	0.02921	0.02949
80-84	0.02804	0.86962	0.03867	0.96133	0.00198	0.09839	0.10037
85-89	0.05274	0.77673	0.07768	0.92232	0.00662	0.15728	0.16390
90-95	0.09922	0.67527	0.15605	0.84395	0.01760	0.19032	0.20792
95-100	0.18666	0.67375	0.31350	0.68651	0.02188	0.09583	0.11771

Table 4. Life Table, female population 2011.
Source: Authors' compilation.

According to formula (1) and the estimated life tables, we can obtain the expected present value of the pensions that a recipient could draw. For practical purposes, we assume a married individual aged 65 who is receiving a retirement pension. If this individual is, however, a widow/er they will also be entitled to a widowhood pension. We suppose two possible scenarios (Table 5). In scenario 1 the individual receives the minimum legal pension and in scenario 2 they draw the average pension[5].

	Minimum benefit	Average Pension
Total	166,862.65€	278,936.04€
Male	137,194.9€	257,835.04€
Female	279,253.97€	340,788.97€

Table 5. Expected Present Values by Pensioner.
(Both retirement and widowhood) Source: Authors' compilation.

Note that the cost for women is higher when considering both the minimum benefit and the average pension.

[5] In 2014, the minimum retirement pension in Spain was 600.30€ per month and the minimum widowhood pension was 632.90€ per month. The average retirement pension was 1,123.93€ per month; 1,256.88 (male), 837.87 (female). The average widowhood pension was 656.31€ per month; 469.64 (male), 667.07 (female).

4. CONCLUSION

We can conclude that women generate a higher cost in both retirement and widowhood pensions because they have a longer life expectancy in the two pure statuses and in the transition status.

At the outset, the probability of being married at age *x+1*, p_x^{cc}, is higher for men than it is for women, but this trend changes approximately in the middle of the series and thereafter remains higher for women than it is for men until the end of the series. In the case of widowhood, the probability of being widow/er at age *x+1*, p_x^{vv}, is higher for women than it is for men throughout the series.

Quantifying the expected expenditure on widowhood pensions may be relevant to analyse possible changes to the current system in Spain. For projections, it is important to take into account the increasing entry of women into the labor market.

ACKNOWLEDGMENTS

The study was supported by the Spanish Ministry of Economy and Competitiveness under grant ECO2012-35584.

REFERENCES

Ahn, N., and Felgueroso, F. (2007). *Adecuación de la pensión de viudedad ante el cambio demográfico y socio-económico,* FIPROS.

CECS (2010). "Incertidumbres en torno a las personas mayores". *Informe España 2010: una interpretación de su realidad social,* Centro de Estudios Científicos, 161-220. Madrid, Fundación Encuentro.

Green, P.J., and Silverman, B.W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall/CRC Press, London.

Haberman, S., and Pitacco, E. (1999). A*ctuarial Models for Disability Insurance.* Chaperman and Hall/CRC Press, London.

Hernández, C.F., Sanabria, M., and Hernández, D. (2011). *La viudedad del futuro. Estudio y propuestas de reforma global para garantizar su sostenibilidad,* FIPROS.

Instituto Nacional de Estadística, INE (2011). *Censo de población y vivienda. Cifras de la Población y Censos Demográficos. Demografía y Población.* Instituto Nacional de Estadística, INE (2011). Defunciones por Comunidad Autónoma de residencia, edad, sexo y estado civil. Estadística de Defunciones. Movimiento Natural de la Población. Fenómenos Demográficos.

Kahale, D. (2011). *La reestructuración de la pensión de viudedad a las nuevas situaciones sociales y las nuevas reformas de familia y convivencia: propuestas de reforma,* FIPROS.

Ministerio de Empleo y Seguridad Social (2002-2013). *Anuario de Estadísticas del Ministerio de Empleo y Seguridad Social.*

Monticone, C., Ruzik, A., and Skiba J. (2008). "Women's Pension Rights and Survivors' Benefits. A Compatative Analysis of EU Member States and Candidate Countries", *ENEPRI Research Report,* 53.

Salvador, C. (2013). "Old-Age Protection for Women in the Spanish Pension System", *International Social Security Review,* 66, 1, 49-70.

Sanabria, A., and Hernández, D. (2010). "La pensión de viudedad. Reforma desde una perspectiva económica a través de la relación entre obligaciones y derechos consumidos", *Economía Española y Protección Social,* 2, 215-238.

3. ON COMPETITIVE NON-LIFE INSURANCE PRICING UNDER INCOMPLETE INFORMATION

H. Albrecher and D. Daily-Amir
University of Lausanne, HEC, Department of Actuarial Science, Switzerland,{hansjoerg.albrecher, dalit.daily-amir}@unil.ch

ABSTRACT

Extending a model of Dutang *et al.* (2013) on a game-theoretic approach for non-life insurance pricing under competition among insurers, we investigate the effects of incomplete information about the price elasticity of competitors. We study Bayesian Nash equilibria and illustrate the sensitivity of equilibrium prices with respect to model assumptions.

1. INTRODUCTION

The pricing of insurance contracts is a classical research topic and several approaches are being used by insurance companies to calculate premiums. Game theory concepts (see e.g. Nash, 1950) have been suggested as a way to evaluate the strategic options available to insurance companies with respect to profitability under competition. The classic reference in this context is Rothschild and Stiglitz (1976) who consider competing insurance firms offering contracts with different premium and deductible. In their model there might be no equilibrium point for this type of competition in the insurance market. Lemaire (1980) applied cooperative game theory concepts for insurance applications and risk-sharing agreements. Other models involving game-theoretic approaches in the insurance context can e.g. be found in Taylor (1986), Rees *et al.* (1999), Emms and Haberman (2005), Polborn (1998) and Powers and Shubik (2006).

The approach adopted in this paper is based on Dutang *et al.* (2013) where a model is proposed to calculate premiums in a non-cooperative game among non-life insurers (see Section 2 for details). Whereas Dutang *et al.* (2013) assume full information among all market participants, in this note we are interested in effects of incomplete information on resulting price equilibria. In particular, we study Bayesian Nash equilibria.

2. ONE PERIOD MODEL AND SOLUTION CONCEPTS

Assume that I insurers compete for a fixed number n of policyholders with homogeneous risks. The policies are issued for one period (say, one year). Every insurer fixes his premium x_j at time 0 within a possible 'action' set A_j. After x_j is chosen by every insurer, each policyholder reacts by either renewing the policy with the present company or switching to a competitor. It is assumed that the insurers offer the same product to all clients so there is no price distinction between renewal policies and new policies, and the insurers do not cooperate to set their prices. In Dutang *et al.* (2013) it was shown how a lapse model following a multinomial logit dynamic motivates the demand function

$$D_j = 1 - \beta_j\left(\frac{x_j}{m_j(x)} - 1\right)$$

of insurer *j*, giving an approximation for the number of renewed policies, where mj $m_j(x) = \frac{1}{I-1}\sum_{j \neq l} x_k$ is the average price of the competitors and $\beta_j > 0$ is a price sensitivity (elasticity) parameter representing the lapse behavior of policyholders relative to company *j*. Each insurer aims to maximize the expected profit defined by

$$\mathcal{O}_j(x) = D_j * (x_j - \pi_j), \qquad (1)$$

where the break-even premium π_j is a weighted average of the actuarial premium of insurer *j* (based on individual claim experience of insurer *j*) and the market premium (based on collective claim experience of the entire market). In addition, one has to ensure that the choice of premium x_j satisfies a solvency constraint imposed by the

regulator as well as possible general premium restrictions, $x_j \in [\underline{x}, \overline{x}]$. It turns out that the typical solvency constraint based on Solvency II considerations and also the usual bounds for the premium are met in any case for the concrete values implemented later, so that we will not consider these constraints further in the present investigation.

We now add the feature of incomplete information to the model, and propose to apply the concept of Bayesian games. In a *Bayesian game,* players do not necessarily know about all the relevant characteristics of the other players; they only have prior beliefs about player *j* to be of a 'type' t_j with probability $p(t_j)$. The typical way to deal with this in game theory is to behave as if one plays against all those types t_j of player *j* at the same time (see e.g. Harsanyi, 1966) and tries to maximize the expected value of the objective function Oi with respect to the distribution of the various types. A so-called Bayesian Nash *equilibrium* (BNE) is defined as the set of actions $x^\star = (x_1^\star, \ldots, x_I^\star)$ that maximize the expected value of the objective function $\mathbb{E}(\mathcal{O}_i)$ calculated with prior probabilities of the players' types $p(t_j)$. Following Harsanyi (1966), in order to calculate the BNE, one should replace each player *j* with $mj > 1$ possible types by m_j players, each with one type and determine $x^\star = (x_1^\star, \ldots, x_I^\star)$ as the combination of premiums satisfying

$$\forall x_j \in A_j : E(\mathcal{O}_j(x_j^\star(t_j), x_{-j}^\star(t_{-j}))) \geq E(\mathcal{O}_j(x_j(t_j), x_{-j}^\star(t_{-j})))$$

for all players j and all types tj . The chosen type of objective function ensures that there exists a unique such BNE x^*, cf. Dutang *et al.* (2013).

3. NUMERICAL IMPLEMENTATION AND RESULTS

We assume now that there are three players and the elasticity parameter β_1 of Insurer 1 is not known to the other players, they instead have probabilities $p(t_{1i})$ at their disposal for Player 1 to have elasticity parameter β_{1i}. If this distribution consists of two possible values, then (2) is comprised of four equations and the optimal solution is obtained by finding the zeroes of their respective derivatives, i.e.

$$
\begin{aligned}
4\beta_{11}x_{11} - (1+\beta_{11})(x_2+x_3) - 2\beta_{11}\pi_1 &= 0 \\
4\beta_{12}x_{12} - (1+\beta_{12})(x_2+x_3) - 2\beta_{12}\pi_1 &= 0 \\
4\beta_2 x_2 - (1+\beta_2)\big[p(t_{11})(x_{11}+x_3) + p(t_{12})(x_{12}+x_3)\big] - 2\beta_2\pi_2 &= 0 \\
4\beta_3 x_3 - (1+\beta_3)\big[p(t_{11})(x_{11}+x_2) + p(t_{12})(x_{12}+x_2)\big] - 2\beta_3\pi_3 &= 0.
\end{aligned}
$$

In general, M possible values for β_1 leads to M types for Player 1 and correspondingly $M+2$ equations. The values of the break-even premium $\pi_j = w_j\overline{a}_{j,0} + (1-w_j)\overline{m}_0$ are now chosen according to the loss model and the respective individual claim experience for this example we assume $\bar{a}_{j,0} = (1.1, 1.15, 1.05)$ and $\overline{m}_0$ = 1.1 as well as $w_j = (1/3, 1/3, 1/3)$, i.e.

$\pi_j = (1.1, 1.167, 1.083)$.

In Figure 1 we compare the case where $\beta_1 \in \{1, 5\}$ with probability 1/2 each with the case of four types $\beta_1 \in \{0.75, 1.25, 4.75, 5.25\}$ with probability 1/4 each (with $\beta_2 = 3.8, \beta_3 = 4.6$ throughout). Note that in both cases $E(\beta_i) = 3$ and $Var(\beta_1) = 4$. One notes that more types for the same player (that is more uncertainty about him) leads to increased equilibrium premiums for all players.

In order to illustrate the influence of the number of players with unknown β_j as well as the value of $Var(\beta_i)$, in Figure 2 we give the resulting premium for Player 3 for five different situations:

- full information game with $\beta 1 = 3$,
- two possible types for Player 1 with $\beta_1 \in \{2.75, 3.25\}$,
- two possible types for Player 1 with $\beta_1 \in \{1, 5\}$,
- two possible types for Player 1 and Player 2 with $\beta_1 \in \{2.75, 3.25\}$, $\beta_2 \in \{3, 4.6\}$ (which maintains the expected value for the previously constant value of β_2),
- two possible types for Player 1 and Player 2 with $\beta_1 \in \{1, 5\}$ and $\beta_2 \in \{1.8, 5.8\}$ values.

It turns out that the variance of the price sensitivity has a major influence on the BNE premiums, whereas the fact that additional uncertainty spread over several opponents (for Player 3) changes the respective equilibrium premium only sightly.

It is also of interest to see to what extent it is an avantage or disadvantage in this setup for one particular player to not communicate transparently his β-value. To that end, let us compare the premium of Player 1 (having elasticity parameter $\beta_1 = 3$) in the following three situations:

- full information game with $\beta_1 = 3$ (i.e. his elasticity parameter is available to the opponents)
- Bayesian game with three types $\beta_1 \in \{1, 3, 5\}$
- Bayesian game with three types $\beta_1 \in \{0.55, 3, 5.45\}$.

Figure 3 indicates that Player 2 and Player 3 set higher premiums as the uncertainty about Player 1 increases. In turn, also Player 1 (being of type $\beta_1 = 3$) has to increase his premium in equilibrium, reacting to the other two players' premium decisions. The resulting objective function values for each player are presented in Table 1. It suggests that it can be an advantage not to disclose one's elasticity to the opponents, as the value of the objective function increases with additional uncertainty (in relative terms the increase is slightly larger for Player 1 than for his opponents).

	Var(β_1)=0	Var(β_1)=2.67	Var(β_1)=4
Player 1 (β_1 = 3)	0.397	0.468	0.570
Player 2	0.391	0.451	0.527
Player 3	0.454	0.523	0.611

Table 1. Objective function values

Up to now uncertainty was about the elasticity parameter β of the players. Let us finally look at the situation where Player 1 thinks of deliberately reducing the used value of the break-even premium to, say, 80% of its true value (e.g. in order to gain

market share). The question is whether this information should be passed on to competitors. Figure 4 gives equilibrium premiums (and Table 2 the respective objective function values) for the following situations:

- reducing π_1 and disclosing this information to the competitors (left column in Fig. 4): this leads as expected to a substantially reduced premium and objective function value.

- incomplete information, where the competitors only know that either Player 1 uses the true break-even premium (probability 0.5) or the reduced one (probability 0.5). If Player 1 indeed uses the true (higher) π_1, the resulting equilibrium premium as well as the objective function is lower than for complete information and it is better to communicate this to the competitors. However, if Player 1 uses the smaller π_1, then we see from Table 2 that Player 1 has an incentive not to tell the competitors about it (in other words, creating uncertainty can be an advantage). For the other players, even if Player 1 uses the smaller value, not knowing this with certainty is an advantage in terms of their objective function value, and they will also use higher premiums in that case.

	π_1=1.1	$\pi_1 \in \{0.88, 1.1\}$	π_1=0.88
Player 1	0.397		0.307
Player 1 (π_1 = 1.1)		0.364	
Player 1 (π_1 = 0.88)		0.339	
Player 2	0.391	0.337	0.283
Player 3	0.454	0.392	0.330

Table 2. Objective function values

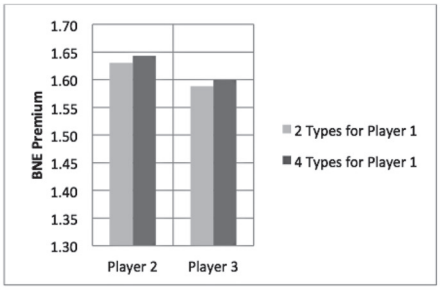


Figure 1. BNE Premiums with 2 or 4 types for Player 1.

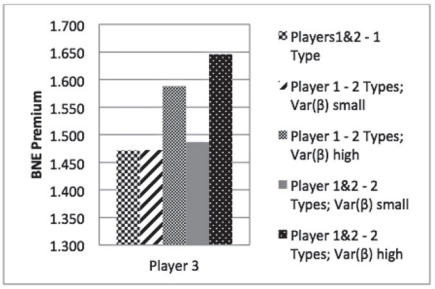


Figure 2. BNE Premiums for Player 3. Various games.

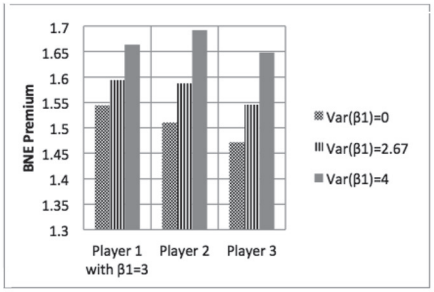


Figure 3. BNE Premiums with varying degree of uncertainty about Player 1.

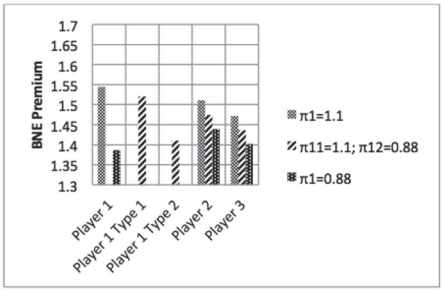


Figure 4. BNE Premiums for two possible π_1.

4. CONCLUSION

In this short note we deal with an extension to the game-theoretic model of Dutang *et al.* (2013) for determining premiums under competition. We focus on the case where not all market participants have full information about the others, which seems to be quite a natural assumption. Studying Bayesian Nash equilibria, we numerically illustrate some scenarios in which uncertainty can be beneficial. The results indicate that uncertainty in a Bayesian game is typically compensated by increased premiums and, depending on the situation, players do or do not have incentives to share information about themselves with their opponents. The obtained quantitative results support the intuition one may have on the effects of incomplete information. Naturally, these findings rely, however, on the particular choice of model, objective function and equilibrium concept. It is left for future research to study similar questions for other choices of all these.

REFERENCES

Dutang, C., Albrecher, H., and Loisel, S. (2013). "Competition among Nonlife Insurers under Solvency Constraints: A Game-theoretic Approach". *European Journal of Operations Research,* 231(3), 702-711.

Emms, P., and Haberman, H. (2005). "Pricing General Insurance Using Optimal Control Theory". *ASTIN Bulletin,* 35(2), 427-453.

Harsanyi, J.C. (1966). "A General Theory of Rational Behaviour in Game Situations". *Econometrica,* 34(3), 613-634.

Lemaire, J. (1980). "Applications of Game Theory to the Insurance Business". *Record of Society of Actuaries*, 6(3), 959-962.

Nash, J.F. (1950). "Non-cooperative Games". *Annals of Mathematics,* 54(2), 286-295.

Polborn, M.K. (1998). "A Model of an Oligopoly in an Insurance Market". *The Geneva Paper on Risk and Insurance Theory,* 23(1), 41-48.

Powers, M.R., and Shubik, M. (2006). "A Square-root Rule for Reinsurance". *Revista de Contabilidade e Finanças (Review of Accounting and Finance),* 17(5), 101-107.

Rees, R., Gravelle, H., and Wambach, A. (1999). "Regulation of Insurance Markets". *The Geneva Paper on Risk and Insurance Theory,* 24(1), 55-68.

Rothschild, M., and Stiglitz, J. (1976). "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information". *The Quarterly Journal of Economics*, 90(4), 629-649.

Taylor, G. (1986). "Underwriting Strategy in a Competitive Insurance Environment". *Insurance: Mathematics and Economics*, 5(1), 59-77.

4. EL PERFIL DEL CONDUCTOR EN ESTADO DE ALCOHOLEMIA: DIFERENCIAS ENTRE CARRETERA Y ZONA URBANA

M. Alcañiz, M. Santolino
Riskcenter-IREA, Universitat de Barcelona

Ll. Ramon
Servei Català de Trànsit,
Generalitat de Catalunya

RESUMEN

Conocer las características de los conductores que circulan bajo los efectos del alcohol es esencial en el diseño de políticas de seguridad vial. En este estudio se identifica el perfil de los conductores que dieron tasas por encima de los límites legales en las pruebas de alcoholemia realizadas en Cataluña (España) en el año 2013. Se analizan las diferencias según la vía de circulación sea interurbana o urbana. Para cada tipo de vía se aplica un modelo de regresión logística para estimar la influencia de diferentes factores sobre la probabilidad de que un individuo conduzca con un nivel de alcohol superior al legalmente permitido. Entre otros resultados, la edad, la nacionalidad, el momento del día o el tipo de vehículo influyen sobre la probabilidad de alcoholemia, y lo hacen de diferente modo, según la circulación sea en vía urbana o en carretera. Estos resultados pueden ayudar a diseñar políticas de seguridad vial más adecuadas para cada tipo de vía.

1. INTRODUCCIÓN

Los accidentes tienen un enorme impacto social y constituyen uno de los mayores problemas de salud pública a nivel mundial. Los costes económicos que se derivan de las lesiones causadas por los accidentes de tráfico representan el 2% del

PIB en los países desarrollados (OMS, 2011). Conducir bajo los efectos del alcohol es uno de los principales factores de riesgo de accidente en la red viaria. En el informe del Instituto Nacional de Toxicología y Ciencias Forenses se estima que casi la mitad de los conductores fallecidos en 2012 había consumido sustancias psicoactivas, siendo el alcohol la sustancia tóxica con mayor presencia, detectándose en 3 de cada 4 casos positivos (INTCF, 2012). Estos resultados están en consonancia con los del proyecto DRUID (Isalberti *et al*., 2011) y otros trabajos previos (Holmgren *et al*., 2005; Wagenaar *et al*., 1995; Ahlm *et al*., 2009).

En este estudio analizamos el perfil de los conductores que conducen bajo los efectos del alcohol en la red viaria catalana. Nuestro objetivo es descubrir si existen diferencias en el perfil de conductor que conduce bajo los efectos del alcohol según el tipo de vía por la que circule: interurbana o urbana. Los trayectos en automóvil realizados en área urbana son frecuentemente cortos. En cambio, es más común que los trayectos llevados a cabo en vías interurbanas sean desplazamientos de media-larga distancia.

En la primera parte del estudio analizamos las diferencias entre el porcentaje de conductores con alcoholemia detectados en vías interurbanas y vías urbanas según sus características demográficas (edad, género, nacionalidad, etc.) y el motivo de realización del control (preventivo, accidente, etc.). En una segunda parte, para cada tipo de vía, se utiliza una regresión logística para modelizar la probabilidad de que un individuo conduzca con un nivel de alcohol por encima del legalmente permitido (variable dependiente) según un conjunto de características observadas o variables explicativas (factores).

2. DATOS

Se dispone de la base de datos de la totalidad de las pruebas de alcoholemia practicadas a los conductores en la red viaria catalana durante el año 2013. Mayoritariamente, las pruebas en carretera fueron realizadas por los *Mossos d'Esquadra* (cuerpo de policía de la Generalitat de Catalunya), y en zona urbana por las policías municipales. La base de datos consta de un total de 464.134 pruebas, de las cuales el 66% se realizaron en vías interurbanas y el 34% en vías urbanas. Se

dispone de información relativa a las características de los conductores a los que se les practicó la prueba, momento del tiempo en que se llevó a cabo, motivo de realización y resultado de la prueba. Esta base de datos es complementaria a la de controles aleatorios de alcoholemia, que se estudió en Alcañiz *et al*. (2014).

3. RESULTADOS

En la Tabla 1 se muestra el ratio de conductores detectados con nivel de alcohol por encima del legalmente permitido, para cada una de las características de interés. Los resultados se presentan para el conjunto de la red viaria catalana y diferenciando el tipo de vía. Como puede apreciarse, el porcentaje de individuos con tasas de alcohol superiores a las legalmente permitidas es sensiblemente superior en vías urbanas (6,3%) que en interurbanas (4,0%). Para algunos factores analizados, el porcentaje de conductores con alcoholemia que se detectó en vías urbanas duplicaba al observado en vías interurbanas. Por ejemplo, sucede así cuando la edad del conductor se sitúa entre los 35 y los 50 años, cuando el motivo de la realización de la prueba fue por infracción de tráfico o accidente, cuando esta se realizó por la noche o en día laborable o si el vehículo era pesado.

Var.	Etiqueta	Descripción	Ratio de conductores con alcoholemia		
			Total	Vía Inter urbana	Vía urbana
x_1	ag>65	El conductor es mayor de 65 años	0,025	0,025	0,024
	ag50-65	La edad del conductor está entre 50 y 65 años	0,032	0,028	0,040
	ag35-50	La edad del conductor está entre 35 y 50 años	0,037	0,027	0,058
	ag25-35	La edad del conductor está entre 25 y 35 años	0,053	0,043	0,069
	ag<25	El conductor es menor de 25 años	0,074	0,072	0,078
x_2	extranjero	El conductor es de nacionalidad extranjera	0,052	0,044	0,065
	español	El conductor es de nacionalidad española	0,047	0,039	0,062
x_3	hombre	El conductor es hombre	0,048	0,039	0,066
	mujer	La conductora es mujer	0,045	0,042	0,050
x_4	prevent	La prueba se realizó en un control preventivo	0,040	0,037	0,047
	infracc	La prueba se realizó por infracción de tráfico	0,268	0,174	0,306
	sintom	La prueba se realizó por síntomas de alcoholemia	0,545	0,598	0,536
	acciden	La prueba se realizó por accidente	0,088	0,069	0,150
x_5	laborable	La prueba se realizó entre lunes y viernes hasta las 15:00	0,031	0,022	0,048
	finsem	La prueba se realizó de viernes a partir de las 15:00 a domingo	0,065	0,058	0,082
x_6	noche	La prueba se realizó por la noche (de 21:00 a 8:00)	0,067	0,054	0,096
	tarde	La prueba se realizó por la tarde (de 14:00 a 21:00)	0,015	0,014	0,015
	mañana	La prueba se realizó por la mañana (de 8:00 a 14:00)	0,015	0,015	0,014
x_7	vpesado	El vehículo es un vehículo pesado	0,010	0,008	0,020
	vmoto	El vehículo es una motocicleta o ciclomotor	0,069	0,043	0,074
	vturismo	El vehículo es un turismo	0,053	0,047	0,064
y_1	positivo	El conductor sobrepasa el límite legal	0,048	0,040	0,063

Tabla 1. Ratio de conductores con una tasa de alcohol por encima de la legal.

Los resultados de la Tabla 1 no recogen la posible relación existente entre factores. Por este motivo, se realiza un análisis multivariante en el que se ajusta una

regresión logística para estimar la probabilidad de que un conductor circule con un nivel de alcohol superior al legal. Este análisis se realiza por separado para las pruebas de alcoholemia llevadas a cabo en vías interurbanas y en vías urbanas. Los resultados se detallan en la Tabla 2.

		Vía Interurbana			Vía urbana		
Variable	Etiqueta	Coef.	Std.Er.		Coef	Std.Er.	
	Constante	-2,61	0,06	***	-2,57	0,08	***
x_1	ag>65	-0,45	0,09	***	-0,65	0,12	***
	ag50-65	-0,23	0,04	***	-0,24	0,04	***
	ag35-50	-0,30	0,02	***	0,01	0,03	
	ag25-35	-	-		-	-	
	ag<25	0,40	0,02	***	0,02	0,03	
x_2	extranjero	0,25	0,03	***	-0,08	0,03	**
	español	-	-		-	-	
x_3	hombre	0,22	0,02	***	0,34	0,03	***
	mujer	-	-		-	-	
x_4	prevent	-1,20	0,04	***	-1,96	0,05	***
	infracc	0,99	0,07	***	0,20	0,06	***
	sintom	2,84	0,15	***	0,97	0,07	***
	acciden	-	-		-	-	
x_5	laborable	-0,65	0,02	***	-0,37	0,02	***
	finsem	-	-		-	-	
x_6	noche	0,96	0,04	***	1,86	0,06	***
	tarde	-0,24	0,05	***	0,04	0,07	
	mañana	-	-		-	-	
x_7	vpesado	-1,16	0,05	***	-0,63	0,07	***
	vmoto	0,10	0,07		0,23	0,03	***
	vturismo	-	-		-	-	
	AIC	93.369			61.300		
	Deviance	93.339			61.270		
	Null deviance	102.054			73.289		
	Nagelkerke						
	pseudo-R2	0,10			0,18		

Nota. *** 1% nivel significación; ** 5% nivel significación.
Tabla 2. Coeficientes estimados de las regresiones logísticas por tipo de vía.

Algunos resultados merecen ser destacados. En relación a la edad del conductor, en vías urbanas no existen diferencias significativas entre los conductores de 25 a 35 años y los conductores menores de 25 años, ni con los conductores de 35 a 50 años. En cambio, en vías interurbanas, es más probable que los menores de 25 años conduzcan con un nivel de alcohol por encima del legalmente permitido que los individuos que se sitúan entre los 25 y 35 años, y a su vez es más probable que estos últimos muestren tasas superiores a la legal que los conductores entre los 35 y 50 años. En relación a la nacionalidad del individuo, el efecto varía si la prueba de alcoholemia se realizó en vía interurbana o urbana. En las primeras, ser extranjero aumenta la probabilidad de conducir con una tasa del alcohol por encima de la legalmente permitida. En cambio, en las vías urbanas, disminuye. Finalmente, señalar que en las vías interurbanas es menos probable que al conductor se le detecte un nivel de alcohol por encima del legal en las pruebas realizadas por la tarde en comparación con las realizadas por la mañana. En cambio, en vías urbanas, las diferencias no son significativas. La probabilidad de alcoholemia es notablemente superior por las noches en ambos tipos de vías.

	Vía interurbana			Vía urbana		
	Odds ratio	IC inferior	IC superior	Odds ratio	IC inferior	IC superior
ag>65	0,64	0,53	0,76	0,52	0,41	0,66
ag50-65	0,79	0,74	0,85	0,78	0,72	0,85
ag35-50	0,74	0,70	0,78	1,01	0,96	1,06
ag<25	1,49	1,43	1,57	1,02	0,96	1,08
extranjero	1,29	1,22	1,35	0,93	0,87	0,98
hombre	1,24	1,18	1,30	1,41	1,33	1,49
prevent	0,30	0,28	0,32	0,14	0,13	0,16
infracc	2,68	2,32	3,10	1,22	1,08	1,37
sintom	17,09	12,75	22,90	2,63	2,28	3,02
laborable	0,52	0,50	0,54	0,69	0,66	0,72
noche	2,60	2,39	2,83	6,45	5,76	7,22
tarde	0,78	0,71	0,87	1,04	0,91	1,19
vpesado	0,31	0,28	0,35	0,53	0,46	0,61
vmoto	1,10	0,97	1,25	1,26	1,19	1,33

Tabla 3. Odds-ratios e intervalos de confianza al 95%

Las diferencias previamente comentadas se basan en el análisis de los factores que influyen de forma inversa según el tipo de vía (cambio de signo del coeficiente

asociado) o bien que tienen capacidad explicativa para una tipología de vía pero no para la otra (coeficiente significativo en una regresión y no significativo en la otra). Sin embargo, los coeficientes pueden ser significativos y con idéntico signo en ambas regresiones pero diferir en la intensidad del efecto. En la Tabla 3 se muestran los *oddsratios* e intervalos de confianza al 95%. Como puede apreciarse, excepto para los factores asociados a edades superiores a 50 años (*ag>65*, *ag50-65*) y el factor que indica si es un vehículo de dos ruedas (*vmoto*), en el resto de los casos el valor que toman los coeficientes estimados es diferente entre las dos regresiones, con un nivel de confianza del 95%.

4. CONCLUSIÓN

En este estudio se investigan los factores que inciden en la probabilidad de que un individuo circule con una tasa de alcohol superior a la legalmente permitida en las pruebas de alcoholemia practicadas en las vías catalanas, diferenciando el tipo de vía de circulación. Para el conjunto de factores analizados, en la mayoría de los casos el efecto de los factores sobre la probabilidad de conducir bajo los efectos del alcohol es diferente si la prueba de alcoholemia se realizó en vía interurbana o en vía urbana.

Identificar las diferencias en el perfil de conductores que circulan con alcoholemia según el tipo de vía de circulación debe ayudar a diseñar políticas de seguridad vial específicas y más adecuadas para las vías interurbanas y urbanas.

AGRADECIMIENTOS

Especial agradecimiento al Servei Català de Trànsit por cedernos la base de datos del estudio. M. Alcañiz y M. Santolino agradecen el apoyo financiero recibido a través del proyecto ECO2012-35584.

REFERENCIAS

Ahlm, K., Björnstig, U., y Öström, M. (2009). "Alcohol and Drugs in Fatally and Non-fatally Injured Motor Vehicle Drivers in Northern Sweden". *Accident Analysis and Prevention*, 41, 129-136.

Alcañiz, M., Guillén, M., Santolino, M., Sánchez-Moscona, D., Llatje, O., y Ramon, Ll. (2014). "Prevalence of Alcohol-Impaired Drivers Based on Random Breath Tests in a Roadside Survey in Catalonia (Spain)", *Accident Analysis and Prevention*, 65, 131-141.

INTCF (2012). *Víctimas mortales en accidentes de tráfico.* Memoria 2012. Instituto Nacional de Toxicología y Ciencias Forenses, Ministerio de Justicia.

Holmgren, P., Holmgren, A., y Ahlner, J. (2005). "Alcohol and Drugs in Drivers Fatally Injured in Traffic Accidents in Sweden During the Years 2000-2002", *Forensic Science International*, 151, 11-17.

Isalberti, C., Van der Linden, T., Legrand, S.A., Verstraete, A., Bernhoft, I.M., Hels, T., Olesen, M.N., Houwing, S., Houtenbos, M., y Mathijssen, R. (2011). *Prevalence of Alcohol and other Psychoactive Substances in Injured and Killed Drivers.* DRUID Driving under the Influence of Drugs, Alcohol and Medicines, 6th Framework Program, Deliverable 2.2.5. Disponible en: https://biblio.ugent.be/publication/1166367 (consultado 5 de diciembre de 2014).

OMS (2011*). Global status report on alcohol and health. Country Profiles*. Geneva: World Health Organization.

Wagenaar, A.C., Zobeck, T., Williams, G.D., y Hingson, R. (1995). "Methods Used in Studies of Drink-drive Control Efforts: A Meta-analysis of the Literature from 1960 to 1991", *Accident Analysis and Prevention*, 27, 307–316.

5. PRICING LIFE INSURANCE WITH TRIANGULAR INTEREST RATES AND FUZZY RANDOM VARIABLES: SOME ANALYTICAL RESULTS FOR MIXED ENDOWMENTS

Jorge de Andrés-Sánchez
Rovira i Virgili University, Department of Business Administration, Spain. jorge.deandres@urv.net

Laura González-Vila Puchades
Barcelona University, Department of Economic, Financial and Actuarial Mathematics, Spain, lgonzalezv@ub.edu

ABSTRACT

This paper develops closed expressions for the defuzzified fair price and the variance of a mixed endowment when discount interest rate is given by means of a triangular fuzzy number. To do it we follow Andrés-Sánchez and González-Vila (2012, 2014) fuzzy-random approach to life insurance pricing.

1. INTRODUCTION

Andrés-Sánchez and González-Vila (2012, 2014) develop an approach to valuating life annuities and insurances that combines the stochastic approach to life insurance (see Gerber, 1995) under deterministic interest rates) and the quantification of interest rates with FNs (Lemaire, 1990; Ostaszewski, 1993; and Andrés-Sánchez and Treceño, 2003, in a life insurance context; and Cummins and Derrig, 1997, and Derrig and Ostaszewski, 1997, within property-liability insurance). This paper extends those results by considering a triangular fuzzy number (TFN) for the mean interest rate used to price policies. In this paper we obtain closed expressions for the mathematical expectation of the present value of a mixed endowment,

its defuzzified value and for the variance. Our results have a straightforward application to any other life insurance structure.

2. FUZZY DISCOUNT FACTOR

We consider that the financial valuation process is done with an average interest rate throughout the pricing horizon. This is a usual way to estimate technical interest rates in actuarial practice and it is based on Fisher's relationship between the nominal interest rate, real interest rate and anticipated inflation. Lemaire (1990) uses a fuzzy interest rate $\tilde{i}$ in this manner. Other authors like Cummins and Derrig (1997) or Derrig and Ostaszewski (1997) consider that $\tilde{i}$ is a prediction based on the expected profit of the insurer's asset portfolio.

Let be the discount annual rate a FN, $\tilde{i}$, whose α -cuts are $i_\alpha = \left[\underline{i}_\alpha, \bar{i}_\alpha\right]$. From $\tilde{i}$ we can define *the discount factor* for 1 m.u. payable in t years as $\tilde{d}_t = \left(1+\tilde{i}\right)^{-t}$. Considering the arithmetic of FNs and that the discount factor is a decreasing function of the interest rate, the α -cuts of $\tilde{d}_t$ are, $\forall \alpha \in [0,1]$:

$$d_{t_\alpha} = \left[\underline{d_t}_\alpha, \overline{d_t}_\alpha\right] = \left[\left(1+\bar{i}_\alpha\right)^{-t}, \left(1+\underline{i}_\alpha\right)^{-t}\right] \tag{1}$$

So, if the mean interest rate is fixed as the TFN $\tilde{i} = (i^{(1)}, i^{(2)}, i^{(3)})$, the α-cut representation of the discount factor is now:

$$d_{t_\alpha} = \left[\underline{d_t}_\alpha, \overline{d_t}_\alpha\right] = \left[\left(1+i^{(3)} - \left(i^{(3)} - i^{(2)}\right)\alpha\right)^{-t}, \left(1+i^{(1)} + \left(i^{(2)} - i^{(1)}\right)\alpha\right)^{-t}\right] \tag{2}$$

It will be useful to obtain the expected interval defined by Campos and González (1989) of the discount factor. So:

$$e_I\left(\tilde{d}_t\right) = \left[\int_0^1 \underline{d_t}_\alpha \, \mathrm{d}\alpha, \int_0^1 \overline{d_t}_\alpha \, \mathrm{d}\alpha\right] = \tag{3a}$$

$$= \left[\int_0^1 \left(1+i^{(3)} - \left(i^{(3)} - i^{(2)}\right)\alpha\right)^{-t} \mathrm{d}\alpha, \int_0^1 \left(1+i^{(1)} + \left(i^{(2)} - i^{(1)}\right)\alpha\right)^{-t} \mathrm{d}\alpha\right]$$

where, for $t = 1$:

$$e_I\left(\tilde{d}_t\right) = \left[\frac{ln\left(1+i^{(3)}\right) - ln\left(1+i^{(2)}\right)}{i^{(3)} - i^{(2)}}, \frac{ln\left(1+i^{(2)}\right) - ln\left(1+i^{(1)}\right)}{i^{(2)} - i^{(1)}}\right] \tag{3b}$$

and for $t \neq 1$:

$$e_I\left(\tilde{d}_t\right) = \left[\frac{\left(1+i^{(2)}\right)^{-t+1} - \left(1+i^{(3)}\right)^{-t+1}}{(t-1)\left(i^{(3)} - i^{(2)}\right)}, \frac{\left(1+i^{(1)}\right)^{-t+1} - \left(1+i^{(2)}\right)^{-t+1}}{(t-1)\left(i^{(2)} - i^{(1)}\right)}\right] \tag{3c}$$

Of course, from (3a) it comes immediately that the expected value is:

$$e_V\left(\tilde{d}_t; \beta\right) = (1-\beta)\int_0^1 \underline{d_t}_\alpha \mathrm{d}\alpha + \beta\int_0^1 \overline{d_t}_\alpha \mathrm{d}\alpha \tag{4}$$

3. PRICING MIXED ENDOWMENTS

Let us consider the most general case of a n-year mixed endowment. The insured party aged x will receive 1 m.u. at the end of the year of his death if this happens within the next n years and 1 m.u. at n if he survives to that date. Due to the stochastic behaviour of the mortality and fuzziness of interest rate, the *present value of the mixed endowment* associated to a person aged x years is a FRV that Andrés-Sánchez and González-Vila (2014) model as:

outcomes	P
$\tilde{d}_r$	${}_{r-1\|}q_x, r = 1,\ldots,n$
$\tilde{d}_n$	${}_n p_x$

(5)

The FRV $\tilde{A}_{x:\overline{n}|}$ defines, $\forall \alpha \in [0,1]$, the infima and suprema $\underline{A_{x:\overline{n}|}}_\alpha$ and $\overline{A_{x:\overline{n}|}}_\alpha$ whose realizations are, respectively, the lower and upper extremes of α -cuts of the FNs that this FRV takes as their values.

From this infima and suprema RVs the fuzzy mathematical expectation and the crisp variance of the FRV $\underline{A_{x:\overline{n}|}}_{\alpha}$ can be obtained. Concretely, for the mathematical expectation, $\forall \alpha \in [0,1]$, $E\left(\tilde{A}_{x:\overline{n}|}\right)_{\alpha} = \left[\underline{E\left(\tilde{A}_{x:\overline{n}|}\right)}_{\alpha}, \overline{E\left(\tilde{A}_{x:\overline{n}|}\right)}_{\alpha}\right]$:

$$\underline{E\left(\tilde{A}_{x:\overline{n}|}\right)}_{\alpha} = E\left(\underline{A_{x:\overline{n}|}}_{\alpha}\right) = \sum_{r=1}^{n} \underline{d_r}_{\alpha}\, {}_{r-1|}q_x + \underline{d_n}_{\alpha}\, {}_{n}p_x \tag{6a}$$

$$\overline{E\left(\tilde{A}_{x:\overline{n}|}\right)}_{\alpha} = E\left(\overline{A_{x:\overline{n}|}}_{\alpha}\right) = \sum_{r=1}^{n} \overline{d_r}_{\alpha}\, {}_{r-1|}q_x + \overline{d_n}_{\alpha}\, {}_{n}p_x \tag{6b}$$

So, if the mean interest annual rate is fixed as the TFN $\tilde{i} = (i^{(1)}, i^{(2)}, i^{(3)})$, considering (1), (6a) and (6b), the above expressions are transformed in:

$$E\left(\tilde{A}_{x:\overline{n}|}\right)_{\alpha} = \left[\sum_{r=1}^{n}\left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-r}\, {}_{r-1|}q_x + \left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-n}\, {}_{n}p_x,\right. \tag{7}$$
$$\left.\sum_{r=1}^{n}\left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-r}\, {}_{r-1|}q_x + \left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-n}\, {}_{n}p_x\right]$$

whereas:

$$e_I\left(\tilde{E}\left(\tilde{A}_{x:\overline{n}|}\right)\right) = \left[\frac{ln\left(1+i^{(3)}\right)-ln\left(1+i^{(2)}\right)}{i^{(3)}-i^{(2)}}\, {}_{0|}q_x + \sum_{r=2}^{n}\frac{\left(1+i^{(2)}\right)^{-r+1}-\left(1+i^{(3)}\right)^{-r+1}}{(r-1)\left(i^{(3)}-i^{(2)}\right)}\, {}_{r-1|}q_x + \frac{\left(1+i^{(2)}\right)^{-n+1}-\left(1+i^{(3)}\right)^{-n+1}}{(n-1)\left(i^{(3)}-i^{(2)}\right)}\, {}_{n}p_x,\right. \tag{8}$$
$$\left.\frac{ln\left(1+i^{(2)}\right)-ln\left(1+i^{(1)}\right)}{i^{(2)}-i^{(1)}}\, {}_{0|}q_x + \sum_{r=2}^{n}\frac{\left(1+i^{(1)}\right)^{-r+1}-\left(1+i^{(2)}\right)^{-r+1}}{(r-1)\left(i^{(2)}-i^{(1)}\right)}\, {}_{r-1|}q_x + \frac{\left(1+i^{(1)}\right)^{-n+1}-\left(1+i^{(2)}\right)^{-n+1}}{(n-1)\left(i^{(2)}-i^{(1)}\right)}\, {}_{n}p_x\right]$$

To obtain the variance of the FRV $\tilde{A}_{x:\overline{n}|}$, we must bearing in mind that

$$V\left(\underline{\tilde{A}_{x:\overline{n}|}}\right)=\frac{1}{2}\int_0^1\left[V\left(\underline{\tilde{A}_{x:\overline{n}|}}_{\alpha}\right)+V\left(\overline{\tilde{A}_{x:\overline{n}|}}_{\alpha}\right)\right]\mathrm{d}\alpha \text{ and that:}$$

$$V\left(\underline{\tilde{A}_{x:\overline{n}|}}_{\alpha}\right)=\sum_{r=1}^{n}\left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-2r}{}_{r-1|}q_x+\left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-2n}{}_{n}p_x-$$

$$-\left[\sum_{r=1}^{n}\left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-r}{}_{r-1|}q_x+\left(1+i^{(3)}-\left(i^{(3)}-i^{(2)}\right)\alpha\right)^{-n}{}_{n}p_x\right]^2 \qquad (9b)$$

$$V\left(\overline{\tilde{A}_{x:\overline{n}|}}_{\alpha}\right)=\sum_{r=1}^{n}\left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-2r}{}_{r-1|}q_x+\left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-2n}{}_{n}p_x-$$

$$-\left[\sum_{r=1}^{n}\left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-r}{}_{r-1|}q_x+\left(1+i^{(1)}+\left(i^{(2)}-i^{(1)}\right)\alpha\right)^{-n}{}_{n}p_x\right]^2 \qquad (9b)$$

then:

$$V\left(\tilde{A}_{x:\overline{n}|}\right)=\frac{1}{2}\left\{\sum_{r=1}^{n}\left(\frac{\left(1+i^{(2)}\right)^{-2r+1}-\left(1+i^{(3)}\right)^{-2r+1}}{i^{(3)}-i^{(2)}}+\frac{\left(1+i^{(1)}\right)^{-2r+1}-\left(1+i^{(2)}\right)^{-2r+1}}{i^{(2)}-i^{(1)}}\right)\frac{{}_{r-1|}q_x}{2r-1}+\right. \qquad (10)$$

$$+\left(\frac{\left(1+i^{(2)}\right)^{-2n+1}-\left(1+i^{(3)}\right)^{-2n+1}}{i^{(3)}-i^{(2)}}+\frac{\left(1+i^{(1)}\right)^{-2n+1}-\left(1+i^{(2)}\right)^{-2n+1}}{i^{(2)}-i^{(1)}}\right)\frac{{}_{n}p_x}{2n-1}-$$

$$-\sum_{r=1}^{n}\sum_{t=1}^{n}\left(\frac{\left(1+i^{(2)}\right)^{-r-t+1}-\left(1+i^{(3)}\right)^{-r-t+1}}{i^{(3)}-i^{(2)}}+\frac{\left(1+i^{(1)}\right)^{-r-t+1}-\left(1+i^{(2)}\right)^{-r-t+1}}{i^{(2)}-i^{(1)}}\right)\frac{{}_{r-1|}q_x\,{}_{t-1|}q_x}{r+t-1}-$$

$$\left.-2\sum_{r=1}^{n}\left(\frac{\left(1+i^{(2)}\right)^{-r-n+1}-\left(1+i^{(3)}\right)^{-r-n+1}}{i^{(3)}-i^{(2)}}+\frac{\left(1+i^{(1)}\right)^{-r-n+1}-\left(1+i^{(2)}\right)^{-r-n+1}}{i^{(2)}-i^{(1)}}\right)\frac{{}_{r-1|}q_x\,{}_{n}p_x}{r+n-1}\right\}$$

5. CONCLUSIONS, EXTENSIONS AND FUTURE RESEARCH

This paper extends the results of Andrés-Sánchez and González-Vila (2012, 2014) to the case of triangular fuzzy technical interest rate, very usual in literature, and endowment contracts. Firstly, it develops closed formulas for the defuzified value of the fuzzy present value, i.e. its expected value and its expected interval. In our opinion those values can be very useful for the actuary when a concrete value of net provisions, pure premiums, etc. must be defined. Likewise this paper also develops an analytical expression for the standard deviation of the insurance present value that may be of interest when quantifying stabilizing reserves or premium recharges.

Finally, let us remark that the developments of this paper can be easily extended to any kind of life insurances and annuities.

REFERENCES

Andrés-Sánchez, J. de, and González-Vila, L. (2012). "Using Fuzzy Random Variables in Life Annuities Pricing". *Fuzzy sets and Systems,* 188, 27-44.

Andrés-Sánchez, J. de, and González-Vila, L. (2014). "Pricing Endowments with Soft Computing". *Economic Computation and economic cybernetics studies research* 1, 124-142.

Andrés-Sánchez J. de, and Terceño, A. (2003). "Applications of Fuzzy Regression in Actuarial Analysis". *Journal of Risk and Insurance,* 70, 665-699.

Buckley, J.J. (1987). "The Fuzzy Mathematics of Finance". *Fuzzy Sets and Systems,* 21, 257-273.

Campos, L.M., and González, A. (1989). "A Subjective Approach for Ranking Fuzzy Numbers". *Fuzzy Sets and Systems*, 29, 145-153.

Cummins, J.D., and Derrig, R.A. (1997). "Fuzzy Financial Pricing of Propertyliability Insurance". *North American Actuarial Journal*, 1, 21-44.

Derrig, R.A., and Ostaszewski, K. (1997). "Managing the Tax Liability of a Property Liability Insurance Company". *Journal of Risk and Insurance*, 64, 695-711.

Feng, Y., Hu, L., and Shu, H. (2001). "The Variance and Covariance of Fuzzy Random Variables and their Application". *Fuzzy Sets and Systems,* 120, 487-497.

Gerber, H.U. (1995). *Life Insurance Mathematics,* Springer-Verlag: Berlin.

Lemaire, J. (1990). "Fuzzy Insurance". *Astin Bulletin,* 20, 33-55.

Li Calzi, M. (1990). "Towards a General Setting for the Fuzzy Mathematics of Finance". *Fuzzy Sets and Systems,* 35, 265-280.

Ostaszewski, K. (1993). *An investigation into possible applications of fuzzy sets methods in actuarial science,* Society of Actuaries: Schaumburg (USA)

6. ALTERNATIVE BOUNDED DISTRIBUTIONS TO MODEL CHINESE FINANCIAL RETURNS

F. Aranda
Phd student, Granada University

C. García and J.M. Pérez
Granada University, Quantitative Methods to economics and business, Granada, Spain, {cbgarcia, josemag}@ugr.es

ABSTRACT

Empirical evidence has shown that the financial returns present left asymmetry and high kurtosis. Another common feature is that the domain of financial returns is limit. However, the normal distribution and another unbounded distributions have been applied to fit this type of data. We propose the Two-Sided Power (van Dorp and Kotz, 2002) and the Uniform-TSP (López *et al.*, 2012) as bounded alternative distributions to model financial returns. We present an empirical application using the data of Hang Seng index from 18 August of 2010 to 8 July of 2014. We note that the coefficient of kurtosis of this data is not so high as was expected which could be motivated by the chinese price regulation. We conclude that the U-TSP distribution presents the best fit and we propose to use this distribution or another bounded distributions to calculate risk measure such as the Value at Risk.

1. INTRODUCTION

Although the Chinese stock market is the second largest stock market behind US stock market, it has some relevant peculiarities that difference this market from developed market stock. Firstly, the most of investors are individual contrarily to developed markets. In addition, the short-sale trading is strictly prohibited and

there is a limit of 10 percent in the daily movement in share price since 16 december 1996, (Dongwei and Fleisher, 1997). Another difference is that we can find different types of shares: A shares restricted to domestic investors and B shares available to both domestic and foreign investors. In addition, H shares are traded on the Hong Kong Stock Exchange in Hong Kong dollars but regulated by Chinese law. The Hang Seng index includes the largest and most liquid stocks incorporated in mainland China that are traded on the Hong Kong Stock Exchange. Wang and Hui (2001) analyzed the statistical properties of the Hang Seng index with minute by minute data from January 3, 1994 to May 28, 1997. They concluded that this index cannot be described by the normal distribution and proposed to use a truncated Levy distribution with an exponential fall-off in its tails. Izzuan and Li (2015) presented a revision about the Chinese stock market and model the distribution of its extreme returns.

The normality assumption presented by Fama (1965) has been displaced due to empirical evidence shows that financial returns are leptokurtic and generally skewed to the left. In addition to light asymmetry and high kurtosis, the financial returns present another relevant feature: they are defined in a limit domain. Despite of this fact, the financial returns are generally fitted with unbounded distribution. In the same way, the traditionally risk measures are obtained from unbounded distributions. It seems evident that if the financial returns are finite, they should be fitted with a bounded distribution. This assumption could be even more logic in the case of Chinese financial returns taking into account the restriction on the stock movement.

In this research, we analyze if the price restriction leads to some peculiar feature that could motivate the use of bounded distribution to fit the Chinese financial returns. In methodology section we revise the Two-Sided Power distribution (van Dorp and Kotz, 2002) and the U-TSP distribution (López *et al.* 2012) and in the empirical application section we compare the behaviour of these distributions with the normal one to fit the Hang Seng index. The last section presents the main conclusions.

2. METHODOLOGY

In the analysis of the behaviour of financial asset returns, a high percentage of empirical distributions present the modal value at zero, revealing little or no asymmetry, and fat

tails. By following Hahn (2008) and using the generating system presented by van Dorp and Kotz (2002), López *et al.* (2012) applied different bounded distributions with a broad range of kurtosis and asymmetry that have the capacity to recover the characteristics of the empirical data. We focus on the family of Elevated Two Sided Power (ETSP) presented by García *et al.* (2011) and, specifically, in the mixture between the Two-sided Power distribution, presented by van Dorp and Kotz (2002), and the Uniform distribution, (U-TSP), by using the parameter λ varying between zero and one.

$$G(t|\theta, P(y|n,\lambda)) = \begin{cases} t\left[\lambda\left(\frac{t}{\theta}\right)^{n-1} + (1-\lambda)\right], & 0 < t \leq \theta; \\ 1-(1-t)\left[\lambda\left(\frac{1-t}{1-\theta}\right) + (1-\lambda)\right], & \theta < t \leq 1\,. \end{cases} \quad (1)$$

Thus, when λ = 0 we recover the uniform distribution and when λ = 1 we obtain the TSP distribution. Note that when the mixture parameter λ takes values close to zero we will observe fat tails as consequence of a greater presence of the uniform distribution.

We will apply the U-TSP distribution to fit the Chinese financial returns and compare the results with the TSP and the normal distributions. We estimate the parameters of the distributions using the least squares methods by minimizing the squared error loss function $\sum_{i=1}^{1000}\left\{G(x_i|\theta) - \frac{i}{1000}\right\}^2$ relative to i=1 1000 the empirical cumulative distribution function (cdf) of the data, where $G(\cdot|\theta)$ is the theoretical distribution to be fitted. In order to analyze the suitability of the adjustments performed, we will use the Sum of Squares Residuals (SSR), the Kolmogorov-Smirnov statistic and the p-value.

3. EMPIRICAL APPLICATION

The Hang Seng Index was launched on 24 November 1969 and it has become the most widely quoted indicator of the performance of the Hong Kong stock market. We are working with 1000 daily historic data from 18 August of 2010 to 8 July of 2014 of the Hang Seng Index obtained from finance.yahoo.com. Returns are obtained as $R_t = \ln\frac{p_t}{p_{t-1}}$. Figure 1 presents the daily log returns where we can note the price regulation.

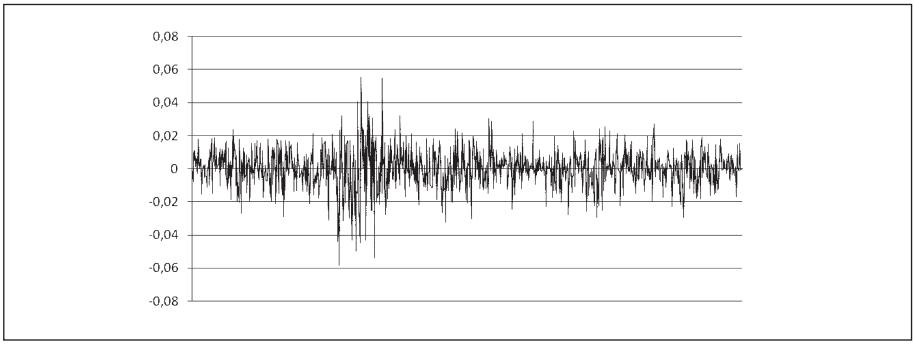


Figure 1. Daily log returns for Hang Seng index from 18 August of 2010 to 8 July of 2014.

Analyzing the returns kurtosis of the fifty companies which composed the index, 16 assets present a coefficient of kurtosis lesser than 3, 23 present a coefficient of kurtosis between 3 and 9 and 11 present a coefficient of kurtosis higher than 9. In relation to asymmetry, 38 asset present left asymmetry and only 12 present right asymmetry. Table 1 displays the mean statistical characteristics of the index. Note that it is asymmetric to the left and with a kurtosis close to three. This coefficient of kurtosis is lower than expected and it could be explained by the strong price regulation in Chinese stock market.

n	Mean	Variance	Symmetry	Kurtosis
1000	0.00015033	0.00013616	-0.28333891	2.85175782

Table 1. Mean statistical characteristics of Hang Seng Index

In Table 2 we present the estimated parameters. Note that the parameter λ in the U-TSP distribution is equal to 0.9373 which means that the introduction of the Uniform distribution is needed to obtain a better fit. On the other hand,

	Distribution	Parameters
Normal	$\mu = 0.516763$	$\sigma = 0.084444$
TSP	$\theta = 0.521077$	$n = 6.053284$
U-TSP	$\theta = 0.520956$	$n = 6.652446\ \lambda = 0.937339$

Table 2. Estimated parameters.

Distribution	SSR	K-S	p-value
Normal	0.418923501	0.037309806	0.123547209
TSP	0.080496001	0.023369899	0.645670449
U-TSP	0.058522245	0.024055630	0.609176530

Table 3. Fit goodness measures.

Table 3 displays the fit goodness measures. We appreciate that the Normal, the TSP and the U-TSP distributions overpass the Kolmogorov-Smirnov Test and the one which presents the lowest Sum of Squares Residuals (SSR) is the U-TSP.

4. CONCLUSION

We have analyzed the Hang Seng index from 18 August of 2010 to 8 July of 2014. As expected, its mean is close to zero and it is asymmetric to the left. However, its coefficient of kurtosis of 2.8 is lesser than expected. This fact could be motivated by the regulation price. If we compare the fit of the normal, TSP and U-TSP distributions we can conclude that the UTSP presents the best fit with the lowest Sum of Squared Residual (SSR). We consider that if the returns are better fitted with a heavy-tail bounded distribution, it could be logical to use this kind of distribution to measure the risk. We are working in the application of this kind of distribution to calculate the Value at Risk (VaR) and other similar measures.

REFERENCES

Dongwei, S., and Fleisher, B.M. (1997). "Risk, Return and Regulation in Chinese Stock Market". *Journal of Economics and Business,* 50(3), 239-256.

Fama, E. (1965). "The Behaviour of Stock-Market Prices". *Journal of Business*, 38, 34-105.

García, C., García, J., and Van Dorp, J.R. (2011). "Modeling Heavy-tailed, Skewed and Peaked Uncertainty Phenomea with Bounded Support". *Statistical, Methods and Applications,* 20, 463-486.

Hahn, E.D. (2008). "Mixture Densities for Project Management Activity Times: A Robust Approach to PERT". *European Journal of Operation Research,* 188, 450-459.

Izzuan, S., and Li, S. (2015). "Modeling the Distribution of Extreme Returns in the Chinese Stock Market". *Journal of International Financial Markets, Institutions and Money,* 34, 263-276.

López, M.M., García, C.B., and García, J. (2012). "Treatment of Kurtosis in Financial Markets", *Physica* A, 391, 2032-2045.

Van Dorp, J. R., and Kotz, S. (2002). "A Novel Extension of the Triangular Distribution and its Parameter Estimation", *Statistician* 551(1), 63-79.

Wang, B.H., and Hui, P.M. (2001). "The Distribution and Scaling of Fluctuations for Hang Seng Index in Hong Kong Stock Market". *The European Physical Journal B-Condensed Matter and Complex Systems,* 20(4), 573-579.

7. COMPARISON OF RESERVING AND SPATIAL MODELS IN INSURANCE

M. Arató and L. Martinek
Institute of Mathematics, Eötvös Loránd University, Budapest, Hungary {arato, martinek}@cs.elte.hu

ABSTRACT

The appropriate estimation of incurred but not reported (IBNR) reserves is traditionally one of the most important task for property and casualty actuaries. As certain claims are reported many years after their occurrence, the amount and appropriateness of the reserves has a strong effect on the results of the institution. In recent years, stochastic reserving methods had become increasingly widespread, supported by a wide actuarial literature, describing development models and evaluation techniques.

A relevant issue of the premium calculation in insurances is the identification of the sources of risk. The more we can determine the origin of risk the more information we get on the distribution of the number and size of occurring claims. The cardinal aim of the present paper is the comparison of appropriateness of several stochastic estimation methods, supposing different distributional underlying development models. Stochastic reserving is being taken into account as a stochastic forecast, thus the comparison techniques developed for stochastic forecasts can be applied, furthermore, expected to be more informative than the classical mean square error of prediction measure. Individual data sets have been utilized, having applied cross-validation and probabilistic scores to obtain new comparisons among Bayesian and non-Bayesian spatial and non-spatial models.

1. INTRODUCTION

The appropriate estimation of incurred but not reported (IBNR) and reported but not settled (RBNS) outstanding claims is crucial to preserve the solvency of insurance institutions, especially in casualty and property insurance.

Data are usually represented by so called run-off triangles. These are $n \times m$ matrices, consisting of $X_{i,j}$ $(i, j = 1, 2, \ldots)$ claim amounts incurred in year i, and paid with a delay of $j - 1$ years. Elements may also indicate aggregate values, and assume without loss of generality that $n = m$.

Unknown elements for indices $i + j > n + 1$ have to be predicted. Let $X_{i,j}$ denote an incremental value in the triangle, hence the outstanding claims are $\sum_{1<i,j<n} X_{i,j} - \sum_{i+j<n+1} X_{i,j}$. Define ultimate claim value $(U\,C)$ as the sum of observed payments (upper triangle) and outstanding claims (lower triangle). Furthermore, let $C_{i,j} := X_{i,1} + \ldots + X_{i,j}$ denote the aggregate claim values. For extensive descriptions of stochastic reserving see [3, 8]. Most of the stochastic reserving methods estimate the distribution of outstanding claims, hence they can be considered as probabilistic forecasts. Present research focusses on the probabilistic scores as comparison tool of forecasts, see [5], demonstrating that these provide an adequate way of comparing stochastic reserving methods, and expected to be more informative than the classical mean square error of prediction measure or quantile values. A test has been run on a real world data set from an insurance company, generating 2000 scenarios with random draw from the claims, and determining the most appropriate stochastic reserving methods.

With respect to the underlying distributions the following models have been taken into account: Log-normal, Negative Binomial, Over-dispersed Poisson and Gamma model. Considering the models above, the parameters (or the adjusted Pearson residuals) are estimated from the upper triangle, followed by the approximation of the distribution of the lower triangle, conditional on the estimated parameters. In general, it is impossible to provide an analytical description of the underlying distribution. However, it might be estimated via Monte Carlo method, generating 5000 lower triangles, for instance. Furthermore, an additional semi-stochastic method presented in [4] has been applied.

In recent years, the development and application of Markov chain Monte Carlo (MCMC) methods have provided a possibility to characterize and to fit to increasingly large classes of spatial models. In practical implementations, a decision has to be made in the final step regarding the model applied. On the basis of the Bayesian models it is feasible to apply Bayesian information criterion or deviance information criterion in this setting and in order to compare the models. Instead of using only classified general databases, one can analyze individual data sets, as well, to investigate new comparisons among different models. We now are now aware of the policyholder's exact age, the zip code of residence, the power of the car the insurance is linked to, etc. Hence, the number of categories is often very high and it produces a new difficulty and a challenge for actuaries.

2. PROBABILISTIC FORECASTS

Most applications of probabilistic forecast theory is used in meteorology. [6] gives an overview of the most important methods regarding the forecasts. Consider the following simple example.

Example 2.1. *Suppose that an automobile insurance company already has paid 1 million euros for the accidents occurred in year 2012. Let ξ be the total amount paid for accidents of 2012 till the end of 2013. Assume that ξ has log-normal distribution with log-scale and shape parameters μ and 1, where μ is supposed to be a standard normal random variable.*

Assume that 4 imaginary actuaries predict the distributions in the above examples, and compare their performances. The ideal actuary knows all the relevant circumstances, which means the knowledge of the exact value of μ. The long-term actuary does not intend to get involved in the actual information, and simply estimates the unconditional distribution. The ordinary actuary attempts to estimate the parameters with some error. At last, the intern actuary finds the expected value, without dealing with the distribution itself. Table 2 provides an overview of the distributions and the predicted distributions.

Actuary	predictive distribution
ideal	$LN(\mu, 1)$
long-term	$LN(0, 2)$
ordinary	$1/2\ LN(\mu, 1) + 1/2\ LN(\mu + \delta, 1)$, where $\delta = \pm 1$ with probability $1/2$, $1/2$
intern	$LN(-\lvert\mu\rvert, \sigma 2)$, where $\sigma 2 = 4\mu + 1$ if $\mu \geq 0$ and 1 otherwise

Table 1. Distributions given by actuaries in the Log-normal example. actuary predictive distribution.

Actuary	Coverage (%)		Average width		CRPS	
	66%	90%	66%	90%	MSEP	
ideal	67.0	90.5	1.93	3.29	-0.48	34.5
long-term	67.3	90.5	2.74	4.65	-1.86	47.2
ordinary	67.2	90.4	5.01	11.64	-0.63	37.2
intern	47.3	74.2	2.87	4.88	-1.83	34.5

Table 2. Coverage, average width, CRPS, MSEP figures (example).

In the evaluation of the predictions, let F_i denote the predicted distribution in the ith case, and x_i a realization. [2] suggested using the Probability Integral Transform (PIT). Prediction is checked by plotting the empirical CDF of the corresponding PIT values and comparing it to the identity function. For PIT histograms regarding the example see url.[1] Long-term, ordinary (and ideal) related predictions do not deviate materially from uniform distribution. Nevertheless, the intern case shows inappropriateness immediately.

Two linked measures are taken into consideration for each probabilistic forecast, based on the idea in [1]. See Table 2 for an example. Coverage in % related to a central prediction interval $\alpha \cdot 100\%$ is the proportion of observations between the lower and upper $\frac{1-\alpha}{2}$ quantiles. Furthermore, the average width for a central prediction interval is the difference of lower and upper $\frac{1-\alpha}{2}$ quantiles expressed in payments, in expectation. It can be interpreted as the sharpness of the prediction. If the performances of two stochastic claims reserving techniques are relatively close to each other, measured in scores, the one with the narrower width is the better.

[1] http://www.math.elte.hu/~arato/stochreserve/stochres_supplement.pdf

Probability scores are devoted to measure the quality of the predicted distributions. A score is a function of the predictive distribution and a realization of the real value. Forecast models are then ranked by comparing the average score of the predictive distributions from each model. *Continuous Ranked Probability Score* (CRPS) can be applied in case of distribution function predictions. Let F denote the predictive distribution function, and cone single observation: $CRPS(F,c) = -\int_{-\infty}^{\infty}(F(y) - 1_{\{y \geq c\}})^2 dy$. The mean square error of prediction (MSEP) of point estimation $\hat{X}$, conditionally on the information of upper triangle is defined as $msep_{X|\mathcal{F}}(\hat{X}) = E((X-\hat{X})^2|\mathcal{F}) = Var(X|\mathcal{F}) + (\hat{X} - E(X|\mathcal{F}))^2$. Table 2 presents the score results in expectation regarding the log-normal example. Furthermore, considering MSEP values, it shows that in the long-term case, the prediction is inaccurate even though the PIT histogram and the prediction interval coverage suggest that the fit is good.

3. SIMULATION AND RESULTS

The goal is to establish a ranking via Monte Carlo simulation among the different stochastic claims reserving techniques, assuming different development models for claims payments, based on scoring rules, instead of MSEP. Suppose that the development distribution model of run-off triangle and corresponding parameters are given.

Step 1. Generate N run-off triangles independently, and besides, for each the ultimate claim values $UC_1, UC_2, \ldots, UC_N$. Let $D_1, D_2, \ldots, D_N$ denote these upper triangles, as conditions to be used later.

Step 2. For each upper triangle, calculate the predicted distribution of the ultimate claims via Monte Carlo method. Assuming these parameter values for each condition upper triangle D_i, ultimate claim values are generated M times, denoted by $\hat{UC}_{i,1}, \hat{UC}^{i,2}, \ldots, \hat{UC}_{i,M}$, as stochastic predictors. The result is a set of predictive distributions. At last, prepare forecast applying the methods described above.

Step 3. For each pair UC_i and tuple $(\hat{UC}_{i,1}, \hat{UC}_{i,2}, \ldots, \hat{UC}_{i,M})$ scores, PIT's, msep and quantile values, and additionally the confidence intervals (coverage and

average width) have been calculated. Generally, a predictive distribution function Fi derives from values $U\hat{C}_{i,1}, U\hat{C}_{i,2}, \ldots, U\hat{C}_{i,M}$, and the corresponding c value is UC_i.

Several question may arise when considering the results: Which score or error number is the most consistent, and how do they correlate with each other? Do the equipments applied in stochastic predictions choose the actual models better than regular measures, such as mean square error of prediction, for instance? Which non-parametric, distributional free methods predict the distributions properly? Do they outperform prediction methods derived from parametric models? How reliable and sharp are the prediction intervals?

A detailed dataset provided by an insurance company (referred as Public Data) contributes to the following example. Figure available external link referred previously shows that in the light of PIT values, none of the methods based on distributional models are recommended. Best histograms are resulted by the bootstrap estimation methods. However, uniform and uniform normal methods are not much worse, they underperform in the cases of extreme claim values. It is also confirmed by the table available on the external link, i.e., the proportion of prediction intervals of bootstrap methods are close to the 66.67% and 90% values. Note that in exchange for the appropriate coverage percentages, average width values are higher.

On the one hand, according to the CRPS values on the referred url, the uniform claims reserving method has the highest score, which has not been performing well for other triangles. It has to be mentioned that the msep values are the highest in the bootstrap cases, in other words, this measure implies a different ranking.

4. SPATIAL DATA

Involving data from more than half a million contracts, an example from nonlife insurance has been analyzed. Special attention has been paid to the simultaneous estimation of the spatial factors and other types of factors. Direct estimation regarding all the 3,111 localities of Hungary is not viable, except for the capital Budapest, due to the lack of policies in certain regions. The occurrence of a single claim in

these sparsely contracted towns results a dramatic increase in the direct risk estimation.

Having supposed that the counting process of claims follows a conditionally independent Poisson model, 4 different approaches has been applied: generalized linear model without spatial dependence, non-spatial credibility model, non-Bayesian spatial smoothing and BYM model. For model selection, the following cross-validation method has been implemented 100 times. 2,000 localities were chosen in the first step, followed by sampling the 70% of policyholders from these towns as a second step. Data obtained from the policies drawn are only used for parameter estimation according to different methods. Having estimated these parameters, the claim numbers have been predicted for all "missing" policies. We propose some different scores between real and predicted data and we analyze their properties.

5. CONCLUSIONS

In general, scores provide a reasonable fit to the background distribution. However, the methods used in this paper and in insurance industry result no reliable prediction intervals. This negative result is not surprising, since a typical run-off triangle contains much less information than the unknown parameters. To improve the predictions, [7] proposed the usage of Bayesian methods. In our opinion, other than these methods, it would also be worthwhile to try using individual contract and claim data in probabilistic forecasts. An individual claim dataset of an insurance company has been used, but not for probabilistic forecasting, but to propose a technique for comparing stochastic methods. There is still room for improvement regarding the sufficient mathematical establishment of this technique, although may be applied in practice. One can simulate many scenarios from the past, and choose a model, which best fits the claims data of the insurance institution.

Both in theoretical and in applied situations, it is worthwhile to test the quality of the different methods, and in multiple ways if possible. The goal was not the construction of the best stochastic claims reserving technique, but to propose an adequate methodology for comparisons in the future. We do believe to have taken the first step in this direction with our work.

REFERENCES

[1] Baran, S., Nemoda, D., and Horányi, A. (2013). "Statistical Post-processing of Probabilistic Wind Speed Forecasting in Hungary". *Meteorologische Zeitschrift,* 22(3): 273-282.

[2] Dawid, A.P. (1984). "Present Position and potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach". J*ournal of the Royal Statistical Society. Series A (General),* 147(2): p. 278-292.

[3] England, P.D., and Verrall, R.J. (2002). " Stochastic Claims Reserving in General Insurance". *British Actuarial Journal,* 8(3): 443-518.

[4] Faluközy, T., Vitéz, I.I., and Arató, M. (2007). *Stochastic Models for Claims Reserving in Insurance Business,* chapter 13, pages 102-113. World Scientific.

[5] Gneiting, T., and Raftery, A.E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". *Journal of the American Statistical Association,* 102 (477): 359-378, March.

[6] Gneiting, T., Balabdaoui, F., and Raftery, A.E. (2007). "Probabilistic Forecasts, Calibration and Sharpness". *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 69(2): 243–268, 2007. ISSN 1467-9868.

[7] Meyers, G. (2007). "Thinking Outside the Triangle", paper presented to the 37th Astin Colloquium, Florida.

[8] Wüthrich, M.V., and Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance.* Wiley.

8. MODELING BIVARIATE KUMARASWAMY DEPENDENT RISKS

B.C. Arnold
University of California, Riverside, Department of Statistics, USA,
barry.arnold@ucr.edu

I. Ghosh
Austin Peay State University, Clarksville, TN, Department of Mathematics and Statistics, USA, ghoshi@apsu.edu

ABSTRACT

A spectrum of bivariate models with Kumaraswamy component distributions is presented. Such models may find application in the study of bounded dependent risks. A variety of construction approaches are considered including copula based models, dependent minima models, conditionally specified models, conditional survival models and transformed bivariate beta models. Multivariate extensions are available for each of the bivariate models discussed.

1. INTRODUCTION

In the context of bounded dependent risks, it is desirable to have available flexible models with analytic expressions for the corresponding marginal distributions and densities. Alternatively, conditional specification might be a convenient approach to the development of suitable models. In such a context, after rescaling the bounded risks to the interval (0, 1), Kumaraswamy distributions provide attractive candidate components for such models because of the simplicity of their corresponding density and distribution functions. It thus becomes of interest to develop a spectrum of bivariate and multivariate models with Kumaraswamy marginal and/or conditional distributions. Our focus in this presentation is on

the bivariate case, but multivariate extensions are readily envisioned for all of the models discussed.

Note that, in the case of unbounded risks, it might be useful on occasions to transform the risks to take on values in the interval (0, 1) in order to take advantage of the flexible arrays of multivariate models that are developed below.

2. THE UNIVARIATE KUMARASWAMY DISTRIBUTION

The Kumaraswamy distribution is defined as follows:

Definition 2.1. A random variable X is said to have a Kumaraswamy distribution with positive parameters α and β if its distribution function is of the form

$$\begin{aligned} F_X(x) &= 0, & x < 0, \\ &= 1 - (1 - x^{\alpha})^{\beta}, & 0 \leq x \leq 1, \\ &= 1, & x > 1. \end{aligned} \tag{1}$$

In this case we write $X \sim K(\alpha, \beta)$. The corresponding density function is

$$f_X(x) = \alpha\beta x^{\alpha-1}(1 - x^{\alpha})^{\beta-1} I(0 < x < 1). \tag{2}$$

In the development of multivariate models with Kumaraswamy component distributions, several distributional properties of the family of Kumaraswamy variables that will prove to be useful are described next.

First, we note a property of closure under minimization.

Theorem 2.1. If $X_1, X_2, \ldots, X_k$ are independent random variables with $X_i \sim K(\alpha, \beta_i)$, $i = 1, 2, \ldots, k$ then $Y = \min_{i=1,2,\ldots,k} X_i \sim K(\alpha, \sum_{i=1}^{k} \beta_i)$.

The Kumaraswamy distribution can be viewed as a generalized beta distribution.

Theorem 2.2. If $X \sim K(\alpha, \beta)$ then $X \stackrel{d}{=} Y^{1/\alpha}$ where $Y \sim Beta(1, \beta)$. Equivalently, X admits a representation of the form $Y \stackrel{d}{=} [U_1/(U_1 + U_2)]^{1/\alpha}$ where $U_1 \sim \Gamma(1, 1)$ and $U_2 \sim \Gamma(\beta, 1)$ are independent random variables.

Not only is the Kumaraswamy distribution function available in a simple form, but the same is also true of its quantile function. Making use of the fact that if X has distribution function F_x then $X \stackrel{d}{=} F_X^{-1}(U)$ where $U \sim Uniform(0, 1)$, we have

Theorem 2.3. If $X \sim K(\alpha, \beta)$ then $X \stackrel{d}{=} [1 - (1 - U)^{1/\beta}]^{1/\alpha}$, where $U \sim Uniform(0, 1)$.

3. USING COPULAS

Recall that a copula $C(x_1, x_2)$ is most simply described as a bivariate distribution function with $Uniform(0, 1)$ marginal distributions. A bivariate Kumaraswamy distribution can then be obtained from any copula $C(x_1, x_2)$ by using marginal transformations of the form indicated in Theorem 2.3 above. Thus if (U_1, U_2) has the copula $C(x_1, x_2)$ as its distribution function, then the random vector

$$(X_1, X_2) = ([1 - (1 - U_1)^{1/\beta_1}]^{1/\alpha_1}, [1 - (1 - U_2)^{1/\beta_2}]^{1/\alpha_2})$$

will have a bivariate Kumaraswamy distribution in which $X_i \sim K(\alpha_i, \beta_i)$, $i = 1, 2$.

This approach thus provides us with a plethora of bivariate Kumaraswamy models. However, it is not clear how to sensibly select the particular copula to be used in the construction. This is an important issue, since the selection of $C(x_1, x_2)$ determines the dependence structure of (X_1, X_2). Beyond the fact that the sign of the correlation in $C(x_1, x_2)$ determines the sign of the correlation in the resulting distribution of (X_1, X_2), it is difficult to interpret how the dependence structure in $C(x_1, x_2)$ will be mirrored in the transformed model.

More flexibility can be built in by beginning with a parameterized family of copulas, say $C(x_1, x_2; \underline{\theta})$, which are then marginally transformed to yield bivariate Kumaraswamy models.

4. THE BARRETO-SOUZA AND LEMONTE MODEL

Barreto-Souza and Lemonte (2012) proposed a bivariate Kumaraswamy model making use of the closure under minimization property of the Kumaraswamy distribution (Theorem 2.1). For this model we begin with independent random variables V_0, V_1, V_2 with $V_i \sim K(\alpha, \beta_i)$, $i = 0, 1, 2$. Then define

$$(X_1, X_2) = (\min\{V_0, V_1\}, \min\{V_0, V_2\}). \tag{3}$$

It follows that $X_i \sim K(\alpha, \beta_0 + \beta_i)$, $i = 1, 2$, where $\beta_0, \beta_1, \beta_2 \in (0, \infty)$.

Each joint distribution corresponding to (3) can be viewd as having been obtained by marginal transformations applied to a Marshall-Olkin copula of the form

$$C_\delta(x_1, x_2) = x_1^{1-\delta} x_2^{1-\delta} [\min\{x_1, x_2\}]^\delta, \ 0 < x_1, x_2 < 1, \tag{4}$$

however the choice of δ will depend on $(\beta_0, \beta_1, \beta_2)$, so that this model is distinct from those described in Section 3.

Note that a more flexible model could be obtained by considering

$$(Y_1, Y_2) = (X_1^{1/\delta_1}, X_2^{1/\delta_2})$$

which will have $Yi \sim K(\alpha\delta_i, \beta_0 + \beta_i)$, $i = 1, 2$.

These models have a possibly unattractive feature. If $\mathrm{V}_0 < \min\{\mathrm{V}_1, \mathrm{V}_2\}$,

which occurs with positive probability, then $\mathrm{X}_1 = \mathrm{X}_2$ and $Y_1 = Y_2^{\,\delta 2/\delta 1}$. Consequently, the joint distribution will not be absolutely continuous (unless one uses a carefully chosen dominating measure). The adjustment proposed by Block and Basu (1974) for a bivariate Marshall-Olkin exponential model could be used here also to define a related absolutely continuous joint distribution.

5. THE NADARAJAH, CORDEIRO AND ORTEGA MODEL

Nadarajah, Cordeiro and Ortega (2011) define a Kumaraswamy-G distribution as follows. Let $G(x_1, x_2)$ be an arbitrary absolutely continuous distribution function. Then for positive parameters α and β define

$$F(x_1, x_2) = 1 - (1 - G^{\alpha}(x_1, x_2))^{\beta}. \tag{5}$$

In order for this to be a valid bivariate distribution function, we must choose $\alpha \geq 1$ and $\beta \in (0, 1]$. If we then choose the function G in (5) to be a copula, say $C(x_1, x_2)$, then it is not difficult to verify that the distribution

$$F(x_1, x_2) = 1 - (1 - C^{\alpha}(x_1, x_2))^{\beta} \tag{6}$$

has Kumaraswamy marginals. Thus if (X1, X2) has (6) as its distribution function, then both X1 and X2 have K (α, β) distributions. Since a positive power of a Kumaraswamy variable also has a Kumaraswamy distribution, we can make marginal power transformations to obtain a more flexible model. In this way we arrive at a model of the form

$$F(x_1, x_2; \alpha, \beta, \delta_1, \delta_2) = 1 - (1 - C^{\alpha}(x_1^{\delta_1}, x_2^{\delta_2}))^{\beta}, \tag{7}$$

where $\alpha \geq 1, \delta_1, \delta_2 > 0$ and $\beta \in (0, 1]$. If(X_1, X_2) has (7) as its distribution function, then it is readily verified that $X_i \sim K(\alpha\delta_i, \beta)$, $i = 1, 2$. In this model the X_i's share the same β parameter. This may be a drawback to use of the model in some circumstances. In addition, the choice of a particular copula to be used in this construction may be difficult to justify, as was the case for the copula models introduced in Section 3.

6. A CONDITIONAL SPECIFICATION APPROACH

Arnold, Castillo and Sarabia (1999) argue that it is sometimes more natural to make distributional assumptions about conditional distributions , instead of about marginal distributions. Arnold and Strauss (1991) provide discussion of distributions with conditional in exponential families. The Kumaraswamy distributions do not constitute

an exponential family, but we can relate them to Beta distributions (Theorem 2.2) which do constitute an exponential family. We can then develop a model with Kumaraswamy conditionals in the following manner. Suppose that (V_1, V_2) has a bivariate distribution with Beta*(1, δ)* conditionals, i.e., $V1|V2 = v2 \sim Beta(1, \beta1(v2))$ for each $v2 \in (0, 1)$ and $V2|V1 = v_1 \sim Beta(1, \beta_2(v_1))$ for each $v_1 \in (0, 1)$. It may be shown that the corresponding joint density of (V_1, V_2) must be of the form

$$f_{V_1,V_2}(v_1, v_2) = C(1-v_1)^{\beta_1-1}(1-v_2)^{\beta_2-1} \exp[\beta_3 \log(1-v_1)\log(1-v_2)]\, I_{(0,v_1,v_2<1)}, \qquad (8)$$

where $\beta1, \beta2, \beta3 > 0$ and C is a normalizing constant which is a function of the β's. Now define $(X_1, X_2) = (V_1^{1/\alpha1}, V_2^{1/\alpha2})$ to obtain a bivariate distribution which, while is does not have Kumaraswamy marginals, does have all of its conditional distributions of the Kumaraswamy form. Thus, for each $x_2 \in (0, 1)$,

$$X_1|X_2 = x_2 \sim K(\alpha_1, \beta_1 + \beta_3 \log(1 - x_2^{\alpha_2})),$$

together with an analogous expression for the conditional distribution of X_2 given $X_1 = x_1$.

7. A CONDITIONAL SURVIVAL APPROACH

Instead of considering the conditional distribution of X_1 given a precise value of X_2, we may instead investigate the consequences of assuming a Kumaraswamy form for the conditional distribution of X_1 given $X_2 > x_2$, with an analogous assumption about the conditional distribution of X_2 . Thus we may assume that for each $x_2 \in (0, 1)$, $X_1|\{X_2 > x_2\} \sim K(\alpha1, \beta1(x_2))$, and for each $x1 \in (0, 1)$, $X_2|\{X_1 > x_1\} \in K(\alpha_2, \beta_2(x_1))$, for some functions $\beta_1(x_2)$ and $\beta_2(x1)$. If this is true, then it may be verified that the joint survival function of (X_1, X_2) is necessarily of the form:

$$\overline{F}_{X_1,X_2}(x_1, x_2) = (1 - x_1^{\alpha_1})^{\beta_1}(1 - x_2^{\alpha_2})^{\beta_2} \exp[\beta_3 \log(1-x_1^{\alpha_1})\log(1-x_2^{\alpha_2})]\, I_{(0<x_1,x_2<1)}.$$

This is an attractive bivariate Kumaraswamy distribution since, it not only has Kumaraswamy conditionals but also it has Kumaraswamy marginal distributions. Thus, for it, we have:

$$X_1 \sim K(\alpha_1, \beta_1), \quad X_2 \sim K(\alpha_2, \beta_2),$$

$$X_1|\{X_2 > x_2\} \sim K(\alpha_1, \beta_1 + \beta_3 \log(1 - x_2^{\alpha_2}))$$

and

$$X_2|\{X_1 > x_1\} \sim K(\alpha_2, \beta_2 + \beta_3 \log(1 - x_1^{\alpha_1})).$$

The problem of identifying the class of all conditional survival models with $X_1|\{X_2 > x_2\} \sim K(\alpha_1(x_2), \beta_1(x_2))$ and $X_2|\{X_1 > x1\} \sim K(\alpha_2(x_1), \beta_2(x_1))$ for some functions $\alpha_1(x_2)$, $\beta_1(x_2)$, $\alpha_2(x_1)$ and $\beta_2(x_1)$ is more difficult. It might be conjectured that solutions can only be obtained in cases in which the "α" functions are constants (as above) or if the "β" functions are constants. A similar conjecture can be made regarding models (as in Section 6) which postulate Kumaraswamy distributions for $X_1|X_2 = x_2$ and $X_2|X_1 = x_1$.

8. AN APPROACH INVOLVING FLEXIBLE BIVARIATE BETA DISTRIBUTIONS

Arnold and Ng (2011) describe quite general classes of bivariate Beta random vectors built as bilinear forms in independent Gamma random variables. Included in their discussion are models of the following form which have Beta(1, γ) marginals. Begin with independent random variables $U0$, U_1, U_2 and U_3 with $U0 \sim \Gamma(1, 1)$ and $U_i \sim \Gamma(\beta_i, 1)$, i = 1, 2, 3. Next define (V_1, V_2) by

$$V_1 = \frac{U_0}{U_0 + U_1 + U_3} \text{ and } V_2 = \frac{U_0}{U_0 + U_2 + U_3}.$$

Observe that $V_i \sim Beta(1, \beta_i + \beta_3)$ $i = 1, 2$. Then, as has become customary by this point, define $(X_1, X_2) = (V_1^{1/\alpha 1}, V_2^{1/\alpha 2})$. This random vector (X_1, X_2) has Kumaraswamy marginals, i.e., $X_i \sim K(\alpha_i, \beta_i + \beta_3)$, ß = 1, 2. It may be noted that this model has the same class of marginal distributions as does the Barreto-Souza and Lemonte model (in Section 3), but has a markedly different dependence structure. One potentially disturbing feature of the model introduced in this Section is the fact that the joint density of (X_1, X_2) is not available in a form that does not involve special functions. Nevertheless, effective inference strategies can be implemented. See Arnold and Ng (2011) for discussion of various possible approaches.

9. ENVOI

As promised, a bushel of flexible bivariate Kumaraswamy models have been described. Investigation of the feasibility of identifying distributional features of bivariate risk data sets, which will aid in selection among these models, is an ongoing project.

Also, as promised, in almost all cases natural extensions to k-dimensional settings ($k > 2$) are readily envisioned, albeit with increased notational complexity.

REFERENCES

Arnold, B.C., Castillo, E., and Sarabia, J.M. (1999). *Conditional Specification of Statistical Models.* Springer-Verlag, New York.

Arnold, B.C., and Ng, H.K.T. (2011). "Flexible Bivariate Beta Distributions". *Journal of Multivariate Analysis,* 102, 1194-1202.

Arnold, B.C., and Strauss, D.J. (1991). "Bivariate Distributions with Conditionals in Prescribed Exponential Families". *J. Roy. Statist. Soc. Ser. B,* 53, 365-375.

Barreto-Souza, W., and Lemonte, A.J. (2013). "Bivariate Kumaraswamy Distribution: Properties and a New Method to Generate Bivariate Classes". *Statistics,* 47, 1-22.

Block, H.W., and Basu, A.P. (1974). "A Continuous Bivariate Exponential Extension". *J. Amer. Statist. Assoc.,* 69, 1031-1037.

Nadarajah, S., Cordeiro, G.M., and Ortega, E.M.M. (2011). "General Results for the Kumaraswamy-G Distribution". *Journal of Statistical Computation and Simulation.* DOI: 10.1080/00949655.2011.562504.

9. ON A NEW MULTIVARIATE IFR NOTION FOR RANDOM VECTORS BASED ON THE STANDARD CONSTRUCTION

A. Arriaza-Gómez
University of Cádiz, Department of Statistics and Operations Research, Spain, antoniojesus.arriaza@uca.es

F. Belzunce
University of Murcia, Department of Statistics and Operations Research, Spain, belzunce@um.es

J. Mulero
University of Alicante, Department of Statistics and Operations Research, Spain, julio.mulero@ua.es

A. Suárez-Llorens
University of Cádiz, Department of Statistics and Operations Research, Spain, alfonso.suarez@uca.es

ABSTRACT

A random variable is IFR if, and only if, it can be ordered with an exponential random variable in the univariate convex transform order proposed by Van Zwet (1964). In particular, a random variable Y is said to be IFR if there exists an increasing convex transformation which maps the exponential distribution onto Y. Given two random vectors X and Y and an appropriate transformation called standard construction, we provide a new multivariate stochastic order assuming the convexity of this transformation. In this work, we propose a new multivariate IFR notion for multivariate distributions based on comparisons in this new order with a properly defined exponentially distributed random vector. Properties, applications and illustrations of these concepts are given.

1. INTRODUCTION

Notions of ageing play an important role in reliability theory, survival analysis and other fields. Among other interesting applications, stochastic orders are related to the study of ageing (see, for example, Belzunce and Shaked, 2008a, and Belzunce and Shaked, 2008b). In particular, in all these well known criteria, we are comparing the relative performance of a unit under consideration with that of an exponential distribution. Remember that a unit with an exponential life distribution suffers no ageing. See, e.g., Barlow and Proschan (1975) or Lai and Xie (2006) for an overview.

Given a random variable X with absolutely continuous distribution function F, survival function $\bar{F} = 1 - F$ and density function f, the hazard or failure rate is defined as $r(x) = f(x)/\bar{F}(x)$ for all x, such that $\mathrm{F}(\mathrm{x}) < 1$. The hazard rate describes the process of ageing and can be considered as the rate at which a unit fails when survive up to a fixed time x. In fact, when r(x) is a nondecreasing function in x, it is said that X is increasing in failure rate (shortly, IFR).

On the other hand, given two random variables X with distribution funtion F and Y with generalized quantile funtion G^{-1}, with interval supports (finite or infinite), it holds that $Y =_{st} G^{-1}F(X)$. Van Zwet (1964) defined the univariate convex transform order when this function is convex and, in this case, it is said that X is smaller than Y in the **convex transform order,** denoted by $X \leq_c Y$.

The connection between the convex transform order and the IFR distributions is the following. A random variable X which represents the lifetime of a component or system is IFR [DFR] if, and only if, it is smaller [larger] than an exponential distribution E in the convex transform order, i.e, $X \leq_c [\geq_c] E$. The purpose of this work is to propose a new multivariate convex transform order as a natural generalization of the univariate order proposed by van Zwet (1964) and to use it to introduce a new multivariate IFR notion by comparisons of a random vector with a properly defined exponentially distributed random vector. Some applications and examples with known probabilistic models are also showed.

2. A NEW MULTIVARIATE CONVEX TRANSFORM ORDER

Several attempts have been made in the literature to extend the IFR notion to the multivariate case (see, e.g., Roy, 1994, or Roy, 2002). In the multivariate case is possible to find such function Φ which maps a random vector X onto a random vector $\Phi(\mathrm{X})$ with the same distribution that Y. In particular, this function Φ has been already used to introduce multivariate IFR notions (see Arias-Nicolás *et al.*, 2009, and Arias-Nicolás *et al.*, 2010, for example).

First, we recall the construction of such function Φ, the so-called standard construction or Rosenblatt transformation (the reader can find more information in Arjas and Lehtonen, 1978, O'Brien, 1975, Rosenblatt, 1952, or Rüschendorf, 1981). For two random vectors X and Y, the multivariate distributional transform of X and the *multivariate quantile transform* of Y and , respectively, are defined recursively, for $i = 1, \ldots, n$, as

$$\begin{aligned} D_{\mathbf{X},i}(x_1,\ldots,x_i) &= F_{\left(X_i \mid \bigcap_{j=1}^{i-1} X_j = x_j\right)}(x_i), \\ Q_{\mathbf{Y},i}(u_1,\ldots,u_i) &= G^{-1}_{\left(Y_i \mid \bigcap_{j=1}^{i-1} Y_j = Q_{\mathbf{Y},j}(u_1,\ldots,u_j)\right)}(u_i), \end{aligned} \tag{1}$$

for every $(x_1, \ldots, x_n)$ in the support of X and $(u_1, \ldots, u_n) \in (0, 1)^n$. If we consider the transform

$$\mathbf{\Phi} : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \mathbf{\Phi} = \mathbf{Q_Y} \circ \mathbf{D_X}, \tag{2}$$

defined for every $(x_1, \ldots, x_n)$ in the support of X, it can be seen that $\mathrm{Y} =_{st} \mathrm{Q_Y}(\mathrm{D_X}(\mathrm{X}))$ and, hence, the function Φ maps the random vector X onto Y. Belzunce *et al.* (2015) proposed the following generalization of the univariate convex transform order based on the convexity of Φ.

Definition 2.1. Let X and Y be two n-dimensional random vectors. Then, X is said to be smaller than Y in the multivariate convex transform order, denoted by $\mathrm{X} \leq_{mct} \mathrm{Y}$, if, and only if, the i-th component of Φ, $\Phi_i : \mathrm{R}^i \to \mathrm{R}$, is convex in their support for all $i = 1, \ldots, n$.

Roughly speaking, $\mathrm{X} \leq_{mct} \mathrm{Y}$ means that the transformation Φ that maps X onto Y is a multivariate convex function. Hence, the previous definition is a clear generalization of the concept of relative skewness proposed by van Zwet (1964).

An interesting situation where the mct-order can be applied is when two random vectors have the same dependence structure.

Theorem 2.1. Let $\mathrm{X} = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be two random vectors sharing a common copula. Then, $\mathrm{X} \leq_{mct} \mathrm{Y}$ if, and only if, $Xi \leq_c Yi$ for all $i = 1, \ldots, n$.

The previous theorem is useful when, for example, we are dealing with ordered data as we will show among other examples.

3. A NEW DEFINITION OF MULTIVARIATE IFR [DFR] DISTRIBUTIONS

In this section, we introduce a new multivariate ageing notion as a direct extension of the known univariate characterization by comparing with an exponential random variable. First, we introduce a properly defined exponentially distributed random vector.

Definition 3.1. Given an n-dimensional random vector Y, we will say that Y follows a multivariate conditional exponential distribution with paramenter matrix $\mathrm{A} = (\mathrm{a}_{ij}) \in \mathrm{M}_{n\times n}$ and constant vector $\mathrm{b} \in \mathrm{R}^n$, denoted by $\mathrm{Y} \sim MCE(\mathrm{A}, \mathrm{b})$, if $\mathrm{Y} =_{st} \mathrm{AE} + \mathrm{b}$, where A is lower triangular with strictly positive diagonal elements and $E = (E1, \ldots, En)$ denotes a multivariate random vector with independent and identically distributed (for short, i.i.d.) exponential marginal distributions with common parameter $\lambda = 1$.

Given $\mathrm{Y} \sim MCE(\mathrm{A}, \mathrm{b})$, it is remarkable that all marginal distributions are linear combinations of independent shifted exponential distributions. In fact, the name "conditional exponential" for Y comes from the property that all sequential conditional distributions are shifted exponential distributions. Next, we give the definition of the new multivariate IFR ageing notion.

Definition 3.2. Let X be an absolutely continuous random vector. Then, X is said to be mct-IFR [mct-DFR] if there exists a particular multivariate conditional exponential

distribution, $\mathrm{Y} \sim MCE(\mathrm{A}, \mathrm{b})$, such that $\mathrm{X} \leq_{\mathrm{mct}} [\geq_{\mathrm{mct}}] \mathrm{Y}$. We will call the random vector Y an upper bound exponential distribution associated with X.

Some relationships dealing with this new aging notion are given. For example, we analyze the mct-IFR notion when the random vector have independent components or for random vectors where the components are obtained through linear combinations of the components of another random vector which satisfies certain aging property related with this new notion.

On the other hand, we also study the aging of the sequential conditional distributions associated with an mct-IFR random vector where these conditional distributions are given by

$$\left(X_i \left| \bigcap_{j=1}^{i-1} \{X_j = Q_{\mathbf{X},j}(u_1, \ldots, u_j)\} \right. \right)$$

for $i = 1, \ldots, n$.

Examples regarding with some probabilistic models such as normal distributions or generalized ordered statistics are also shown and studied.

4. A COMPUTATIONAL STUDY FOR BIVARIATE CASE

Finally, we show a computational study for some particular bivariate distributions. For instance, we study graphically the normal distribution, the bivariate Pareto and the conditional beta distributions through the plots of the components of Φ and we provide the R-code used along the work to get the different plots.

ACKNOWLEDGMENTS

Félix Belzunce and Julio Mulero are partially supported by Ministerio de Economía y Competitividad (Spain) under grant MTM2012-34023-FEDER.

REFERENCES

Arias-Nicolás, J.P., Belzunce, F., Núñez-Barrera, O., and Suárez-Llorens, A. (2009). "A Multivariate IFR Notion Based on the Multivariate Dispersive Ordering". *Applied Stochastic Models in Business and Industry* 25, 339-358.

Arias-Nicolás, J.P., Mulero, J., Núñez-Barrera, O., and Suárez-Llorens, A. (2010). "New Aging Properties of the Clayton-Oakes Model Based on Multivariate Dispersion". SORT 34, 79-94.

Arjas, E., and Lehtonen, T. (1978). "Approximating Many Server Queues By Means of Single Server Queues". *Mathematics of Operations Research* 3, 205-223.

Barlow, R.E., and Proschan, F. (1975). *Theory of Reliability and Life Testing.* New York: Holt, Rinehart and Winston, Inc.

Belzunce, F., Mulero, J., Ruiz, J.M., and Suárez-Llorens, A. (2015). *On relative skewness for multivariate distributions.* Universidad de Murcia, submitted.

Belzunce, F., and Shaked, M. (2008a). "Stochastic Orders and Aging Notions", in *Encyclopedia of Statistics in Quality and Reliability,* edited by F. Ruggeri, F. Faltin and R. Kenett. Wiley, London, 1931-1935.

Belzunce, F., and Shaked M. (2008b). "Multivariate Stochastic Orders and Aging" in *Encyclopedia of Statistics in Quality and Reliability,* edited by F. Ruggeri, F. Faltin and R. Kenett. Wiley, London, 1224-1227.

Lai, C.D., and Xie, M. (2006). *Stochastic Ageing and Dependence for Reliability.* New York: Springer.

O'Brien, G.L. (1975). "The Comparison Method for Stochastic Processes". *The Annals of Probability* 3, 80-88.

Rosenblatt, M. (1952). "Remarks on a Multivariate Transformation". *Annals of Mathematical Statistics* 23, 470-472.

Roy, D. (1994). "Classification of Life Distributions in Multivariate Models". *IEEE Transactions on Reliability* 43, 224-229.

Roy, D. (2002). "Classification of Multivariate Life Distributions Based on Partial Ordering". *Probability in Engineering and Information Sciences* 16, 129-137.

Rüschendorf, L. (1981). "Stochastically Ordered Distributions and Monotonicity of the OC-Function of Sequential Probability Ratio Tests". *Mathematische Operationsforschung und Statistik Series Statistics* 12, 327-338 .

Van Zwet, W.R. (1964). *Convex Transformations of Random Variables.* Amsterdam: Mathematical Centre Tracts, 7 Mathematisch Centrum.

10. GENDER DIFFERENCES IN THE EFFECT OF DRIVING PATERNS ON THE RISK OF ACCIDENT IN PAYD INSURANCE

M. Ayuso
University of Barcelona,Department of Econometrics, RiskCenter, Spain, mayuso@ub.edu

A.M. Pérez-Marín
University of Barcelona,Department of Econometrics, RiskCenter, Spain, amperez@ub.edu

ABSTRACT

In *Pay-As-You-Drive (PAYD)* automobile insurance the premium depends on the distance travelled and driving patterns of the policyholder. We carried out a study of approximately 16,000 young drivers that had purchased a PAYD insurance and we analyzed their risk of accident as a function of driving patterns. The role of driving experience is analyzed by comparing novice vs. experienced young drivers. We conclude that driving at night and a higher proportion of speed limit violations reduces the time to the first crash and urban driving reduces the distance traveled to the first crash for both groups. We also conclude that nighttime driving reduces the time to the first accident in the case of women, but not for men. The risk of an accident increases with excessive speed, but the effect of speed is significantly higher for men than it is for women among the more experienced drivers.

1. INTRODUCTION

Pay-As-You-Drive (*PAYD*) automobile insurance is a policy agreement tied to vehicle usage. The permission of the driver for installing a Global Positioning System (GPS) in the vehicle is required. The GPS provides the insurer with detailed information on

the use of the car and the premium is calculated based on this usage. In this way, those who use less the car are going to pay less than those who do more kilometers during the year. Additionally, the driving pattern of the policyholder is also considered in order to calculate the premium. Namely, speed, type of road and part of the day when the car is most frequently used by the driver are taken into account in the rating system, as they have been proved to explain the risk of accident (Langford *et al.*, 2008; Ayuso *et al.*, 2010; Jun *et al.*, 2011).

2. MATERIALS AND METHODS

We conducted a study of 15,940 young drivers who underwrote a PAYD policy in 2009 with a leading Spanish insurance company. On signing the agreement, their driving patterns began to be registered using a GPS system. The follow-up period was concluded on 31 December, 2011. All drivers were under the age of thirty at the time of underwriting the policy. In Table 1 the variable description is presented.

Variable	Label
Sex	Binary variable (= 1 male, = 0 female)
Age	Age of the driver when driving patterns began to be recorded (measured in years)
Age Vehicle	Age of the vehicle when driving patterns began to be recorded (measured in years)
Experience	Driving experience, measured by the time elapsed since obtaining license and the point when driving patterns began to be recorded (measured in years)
Urban	% of urban driving (% of the total kilometers traveled in urban areas)
Night	% of nighttime driving (% of the total kilometers traveled at night)
Speed	% of the total kilometers traveled above the mandatory speed limits
Km/day	Driving intensity (average number of kilometers traveled per day)

Table 1. Variable description.

Our statistical analysis compares two subsets of drivers:

a) Novice young drivers: individuals that had been in possession of their driving licenses for less than a year when their driving patterns began to be recorded

and who were under the age of thirty at that time. The sample includes 2,790 novice drivers (17.5%).

b) Experienced young drivers: individuals that had been in possession of their driving licenses for at least a year when their driving patterns began to be recorded and who were under the age of thirty at that time. The sample includes 13,150 experienced young drivers (82.5%).

Information on the number of days and number of kilometers to the first accident is interval censored, since we only know in which interval of time/accumulated traveled kilometers an accident has occurred (time interval windows are, on average, equal to 157 days and distance interval windows are, on average, equal to 5,023.19 km). The survival probability estimates of the time/distance to the first accident are obtained using Groeneboom and Wellner's (1992) algorithm implemented in the EM-ICM SAS macro. In order to analyze the effect of driving patterns on the time and distance to the first accident we use a Weibull regression model.

3. RESULTS

In our sample there are 56.17% male drivers (50.72% in the case of the novice group and 57.32% in that of the experienced group). Table 2 shows some descriptive statistics.

	Novice		Experienced		All	
	Mean.	S.D	Mean	S.D.	Mean	S.D.
Age	21.83	2.89	23.89	2.62	23.53	2.78
Age vehicle	7.36	4.89	5.88	4.09	6.14	4.28
Experience	0.48	0.32	4.24	2.48	3.58	2.67

Table 2. Descriptive statistics for the age of the drivers (*Age*), the age of the vehicles (*Age vehicle*) and the years of experience (*Experience*)

Table 3 shows the mean distance driven per day, the percentage of urban and night-time driving and the percentage of distance driven at speeds above the limit for novice and experienced drivers. Results are presented separately for male and

female novice and experienced drivers. We observe that experienced drivers, on average, cover more km/day than novice drivers, but travel a lower percentage of kilometers comparatively on urban roads. Experienced young drivers also travel a lower percentage of kilometers at night, on average, but are more likely to violate the speed limits. Additionally, men present riskier driving patterns than women. These differences are statistically significant.

		Novice		Experienced		All	
		Mean.	S.D	Mean	S.D.	Mean	S.D.
All	Km/day	30.40	20.33	33.78	21.14	33.19	21.04
	Urban (%)	30.47	15.36	26.86	14.37	27.49	14.61
	Night (%)	7.48	6.35	7.23	6.17	7.28	6.20
	Speed (%)	6.67	6.62	8.94	8.39	8.54	8.15
Men	Km/day	33.58	22.05	36.34	22.24	35.90	22.23
	Urban (%)	31.14	15.32	27.01	14.13	27.66	14.40
	Night (%)	8.59	6.73	8.15	86.28	8.22	6.36
	Speed (%)	7.29	7.11	9.82	8.79	9.42	8.59
Women	Km/day	27.12	17.81	30.35	19.03	29.72	18.84
	Urban (%)	29.71	15.38	26.67	14.70	27.26	14.88
	Night (%)	6.34	5.71	6.01	65.79	6.07	5.77
	Speed (%)	6.02	6.00	7.76	7.67	7.42	7.40

Table 3. Driving patterns by gender and driving experience. Means and standard deviations are shown.

In Figure 1 the survival curves for the time to first collision for novice versus experienced drivers are represented with 95% confidence intervals. They have been obtained by using the non-parametric maximum likelihood estimation from interval censored data implemented in the *interval* R Package (Fay, 2014). The 95% confidence interval of the probability of accident during the first year for novice drivers is (25.51%, 28.72%) while the one corresponding to experienced drivers is (19.97%, 21.67%). Similarly, Figure 2 shows survival curves for the distance travelled to first crash for novice versus experienced drivers. The 95% confidence interval of the probability of accident during the first 10,000 kilometers travelled by novice drivers is (22.51%, 28.68%) while the one corresponding to experienced drivers is (18.20%, 20.73%).

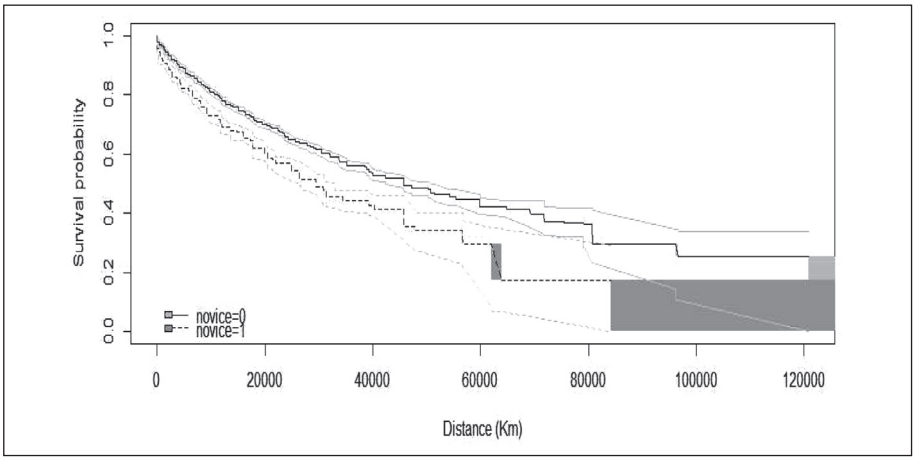


Figure 1. Survival probabilities for the time (days) to the first crash for novice (dashed line) and experienced (solid line) drivers with 95% confidence intervals.

Table 4 shows the results of the Weibull regression model estimations for the time to the first accident and for the distance traveled to the first accident. Both estimations are conducted separately for the group of novice and experienced drivers.

The results were obtained using the Lifereg SAS procedure. When we model the expected time to the first accident, we conclude that the higher the percentage of kilometers traveled above the speed limit, the sooner the accident is likely to occur for both groups of drivers. We also conclude that the driver's gender has an influence on the time and distance to the first accident. Male novice drivers have, on average, their first crash sooner than women. In addition, we find that driving at night reduces the time to the first accident for women, but it has no significant effect for men. Experienced drivers are likely to drive more kilometers until they are involved in their first accident, but excessive speeds reduce the time to the first accident more markedly for men than they do for women.

We report statistical evidence for the effects of other risk factors frequently used in the analysis of accident occurrence. On the one hand, driving experience delays the time to the first accident. In addition, vehicle age has some impact on the risk of accident. We conclude that the older the vehicle, the shorter the distance traveled to the first accident.

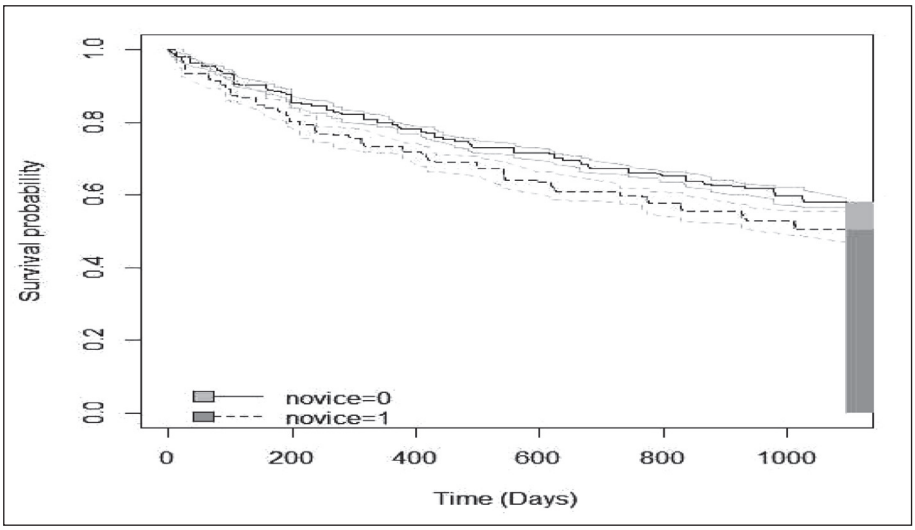


Figure 2. Survival probabilities for the distance (kilometers) travelled to the first crash for novice (dashed line) and experienced (solid line) drivers with 95% confidence intervals.

3. CONCLUSIONS

Our results are of particular relevance for insurance companies offering PAYD insurance contracts. They could be used in insurance ratemaking, as they provide relevant information for creating homogeneous risk classes within the group of young drivers. Additionally, our results are of particular relevance for road safety authorities in their efforts at promoting effective accident prevention programs for specific groups of individuals, such as young drivers.

	Novice young drivers				Experienced young drivers			
	Time		Km		Time		Km	
	Parameter	S. E.	Parameter	S. E.	Parameter	S. E.	Parameter	S. E.
Intercept	7.892[a]	0.105	11.535[a]	0.142	8.159[a]	0.084	12.002[a]	0.083
Sex (men)	-0.309[b]	0.105					0.124[b]	0.029
Urban					-0.030[a]	0.004	-0.056[a]	0.004
Urban^2			-0.038[a]	0.009	$3{\cdot}10^{-4}$ [a]	$1{\cdot}10^{-4}$	$4{\cdot}10^{-4}$ [a]	$1{\cdot}10^{-4}$
Night	-0.033[a]	0.008	$3{\cdot}10^{-4}$ [c]	$1{\cdot}10^{-4}$	-0.012[a]	0.003		
Night *	0.023[c]	0.010						
Sex Speed	-0.032[c]	0.013						
Speed^2	0.001[c]	$5{\cdot}10^{-4}$			-0.030[a]	0.005	-0.012[c]	0.005
Speed *					$7{\cdot}10^{-4}$ [a]	$1{\cdot}10^{-4}$	$3{\cdot}10^{-4}$ [c]	$1{\cdot}10^{-4}$
Sex					-0.006[c]	0.003		
Urban>25%(*)	-0.193[b]	0.066						
Experience					0.044[a]	0.006	0.031[a]	0.006
Age vehicle							-0.017[a]	0.004
Scale	0.797	0.028	0.756	0.026	0.770	0.013	0.890	0.012
Shape	1.255	0.045	1.322	0.045	1.298	0.022	1.124	0.022
-2logL	4220.472		4034.351		19581.09		19389. 69	
AIC	4236.472		4042.351		19599.09		19407. 69	
BIC	4282.965		4065.523		19665.69		19474. 21	

Table 4. Weibull model estimation for the time to the first accident for the distance to the first accident. Results are shown for novice and experienced young drivers separately.

[a] p-value ← 0.0001
[b] p-value ← 0.01
[c] p-value ← 0.05

(*) "Urban → 25%" represents a binary variable that is equal to one when urban driving is higher than 25%.

REFERENCES

Ayuso, M., Guillén, M., and Alcañiz, M. (2010). "The Impact of Traffic Violations on the Estimated Cost of Traffic Accidents with Victims". *Accident Analysis and Prevention*, 42, 709-717.

Fay, M.P. (2014). *Weighted Logrank Tests and NPMLE for interval Censored Data. The Interval R package*, version 1.1-0.1.

Groeneboom, P., and Wellner, J.A. 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation, New York: Birkhauser.

Jun, J., Guensler, R., and Ogle, J. (2011). "Differences in Observed Speed Patterns Between Crash-Involved and Crash-Not-Involved Drivers: Application of In-vehicle Monitoring Technology". *Transportation Research Part* C, 19, 569-578.

Langford, J., Koppel, S., McCarthy, D., and Srinivasan, S. (2008). "In Defence of the 'Low-mileage bias'". *Accident Analysis and Prevention*, 40, 1996-1999.

11. A MODEL FOR RATING TRANSITIONS OVER TIME

S. Baena-Mirabete and P. Puig
Universitat Autònoma de Barcelona, Departament de Matemàtiques, Spain, sergio.baena@uab.cat

ABSTRACT

We propose several parsimonious models for higher-order Markov chains, applied to the study of municipal rating migrations in credit risk. For some processes, as in the case of the credit ratings, the transitions between distant states are unlikely (persistent transitions). We introduce the Short and Long-persistence models and we compare them against the full-parametrized Markov chain achieving a better fit with a less number of parameters. The advantages of these new models are that the results are directly interpretable and they can be fitted using standard statistical packages.

1. INTRODUCTION

The rating agencies, such as Moody's, Standard and Poor's or Fitch, have as their main objective to evaluate the capacity of a debt issuer to pay back its debt as well as the probability that the issuer to fall into default. To this end, the rating agencies provide a risk scale that rates the debt into homogeneous discrete risk classes based on the solvency of the issuer. The rating assigned from these agencies are expressed on an ordinal scale where the position in the scale reflects the ability of the issuer to pay. The rating AAA indicates a strong financial capacity to cope with their debts and the rating C indicates the lowest quality of the firm. Figure 1 shows plots of rating paths over the time for a set of issuers which were drawn randomly from our data basis. The rating classes were recoded to numerical format from 1 to 6, being 1 the lowest rating (CCC) and 6 the highest (AAA).

In recent times, the rating agencies have become the target of financial institutions who feel threatened to downgrade the rating of their debt. That is because of the rating assigned to a debt issuer has direct consequences in its financing capacity and risk management. For instance, the more deteriorated the credit quality of the portfolio of an entity, as a result of the estimated transitions towards lower ratings, the more provisions will be necessary to cover the increment in expected losses.

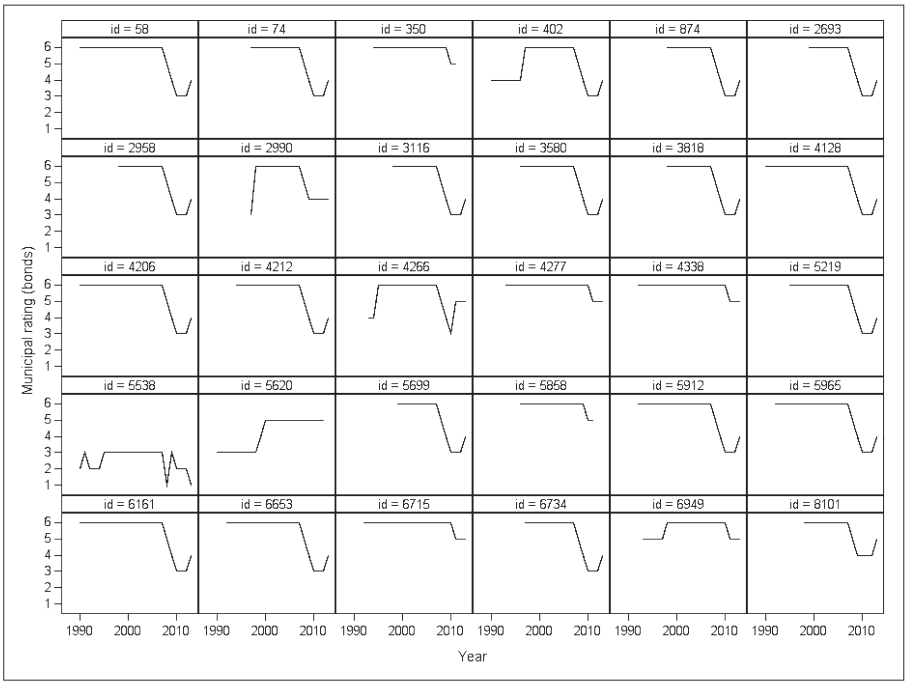


Figure 1. Sequences of ratings over the time for 30 debt issuers randomly selected from the municipal rating sample.

There are numerous empirical studies on the analysis and estimation of rating transitions. However, research and practice in this field have increasingly converged into a general consensus: the knowledge of the previous rating is not sufficient to explain the dynamic of the rating migrations. The simple idea is that under a 1-order Markov chain, two debt issuers with an identical previous rating should have the same probability to be downgraded or upgraded at the current state. But several empirical

studies show just the contrary. An interesting contribution can be seen in Lando and Skødeberg (2002) where they consider the estimation of credit rating transitions based on continuous-time observations. This approach has as main advantage that the estimates are based on the full story of the rating transitions. Basically, two conclusions can be drawn from this study: the first one is known in the literature as "rating momentum" and this means that a previous downgrade increases the probability to be downgraded again. For upgrades, the result is almost the opposite. The second finding of Lando and Skødeberg (2002) concerns the relationship between the time since last entry into the current rating and the transition probability. They found the longer the persistence in a rating the less likely it will be a change of state in the immediate future. In McNeil and Wendin (2006) a class of generalized linear mixed models were fitted for ordered polytomous responses by assuming a Markov process for the random effects. In this model, they took into account implicitly serial correlations in the migration counts that, as the authors pointed out, could be interpreted as momentum effects due to past economic conditions. In Wozabal and Hochreiter (2012), a coupled Markov chain model was proposed to model the rating transitions. However, as noted by the authors, to use this model in practice, it is necessary to implement more sophisticated estimation routines to obtain realistic parameter estimates. The main objective of this article is to analyze the dynamic of the rating migrations as well as to obtain reliable estimates for the probability rating transitions in a one-year time horizon. The data refer to a sample of 4747 municipal bonds issuers in the United States rated by the Credit Rating Agency Moody's for the historical period from 1992 to 2013.

2. THE MODEL

The Markov chain is a well known probabilistic model used to explore dependencies between successive observations. The usual first order Markov chain model considers that, given the present, the future is conditionally independent of the past. However, in some situations the future depends not only on the present but on the last observations. Unfortunately the number of parameters of these fully parameterized models increases very rapidly: a k-order Markov chain model with m states has $m^k\ (m-1)$ parameters. For example, for a Markov chain of order equal to 3 and 6 possible states, the number of parameters to estimate is 1080. From now on MCk

will denote a full parameterized k-order Markov chain model. MC_0 model will correspond to a series of independent outcomes.

A multinomial cumulative (logistic) model can be used to estimate a k-order Markov chain model. However, in the credit data, the transitions between distant ratings are unlikely and the off-diagonal probabilities are poorly estimated due to the low number of observations, occasionally leading us to convergence failures. A common practice is to impose constraints on the parameters. In this regard, the proportional-odds model (POM) described in McCullagh (1980), is highly restrictive and in many practices, an unrealistic assumption. In order to estimate the transition probabilities for the MC_1 model, we consider a less restrictive approach that we call the *Shortpersistence* model (SPM) which only depends on the state at the previous instant. In addition, a *persistent-diagonal* matrix is constructed in order to relax the proportional odds assumption doing use of the persistent nature of the ratings over time.

Furthermore, we propose a more parsimonious higher order Markov model taking into account the long-persistence observed in the ratings migration. The SPM model is extended to that we call the *k-order Long-persistence* model in which includes the immediately previous observations (short persistence) and also a summary of what happened many time before a certain instant (long persistence). Consider a discrete-time random variable X_t taking values in a finite set of states $1, 2, ..., m$. Then, we define the k-order *Long-persistence* Markov Chain model of degree r (from now denoted as $LPM_k(r)$) by means of the relationship:

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) =$$

$$P(X_t = i_t | E_r^k(t), U_r^k(t), D_r^k(t), X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1})$$

where $i_0, i_1, ..., i_t \in \{1, 2, ..., m\}$, k, r are integers and $k > r$. In the conditional part there are the long-persistence terms, $E_r^k(t)$, $U_r^k(t)$, $D_r^k(t)$, that summarize a sequence of states between the instants $t - r - 1$ and $t - k$, and the short-persistence terms represented by the short memory observations $X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1}$.

An advantage of this model is that it is simple to interpret and standard statistical packages can be used for fitting it as the VGLM package in R (Yee, 2010) or the NLMIXED procedure in SAS.

3. THE RESULTS

We have fitted the LPMk (r) models for k=2,3,4,5,6 and r=1,2 to the ratings data. According to the BIC criterium, the best model is the LPM3(1) and this means that the state of the rating at the instant t depends on the 3 previous states. In Table 1 we show a summary of the results of fit for several estimated models: the fully parametrized Markov chain models (MCk), the POM model of order 1, the SPM and the LPM3(1) model. Notice that, MC0, the independence model, is among the worst results. Even though MC0* by adding a random effect by issuer was worse than the higher order Markov chain models explored.

q	BIC	model
5	72520	MC_0
6	67945	MC_0*
30	35592	MC_1
180	33870	MC_2
1080	41525	MC_3
10	36102	MC_1-POM
13	35551	SPM
17	32532	$LPM_3(1)$

Table 1. BIC statistics and number of parameters (q) for several models fitted.

4. CONCLUSION

We have presented a statistical model to fit high order Markov chain models with long-persistence, that is, models for which the transitions typically take place between states which are close. One of the main problems in high order Markov chain models is the high number of parameters to be estimated. However, the long-persistency property of rating transitions has been used to fit more parsimonious Markov chain models in order to estimate the oneyear transition probabilities between states of rating. We have found that in municipal ratings the information of the previous state is insufficient to explain the rating at the present time. Specifically, the probability of observing a particular rating at the current instant depends on the observed states over the last three years. Moreover, we have shown that it is not

necessary to specify all the information concerning the previous history and a summary of this has led us to construct a more simple and explicative model.

Our model can be fitted using a cumulative logistic regression for an ordinal polytomous response, using standard statistical packages. Some times, a common procedure is to impose the proportional odds assumption (POM model) that leads to a lower number of parameters. However, the POM is inappropriate to estimate transitions between states for the municipal ratings sample. A general non-proportional model (NPOM) without constraints on the parameters can fall into convergence problems due to the high number of parameters to be estimated. In persistent rating transitions, this problem is further accentuated by the low frequency of transitions between distant states by leading us to failures in the estimation. Our model is a partial proportional model that becomes less restrictive than the POM, obtaining better results according to the BIC criterium. In our model we take into account a persistent matrix in which the diagonal parameters are not constrained and, in addition, they may be reliably estimated.

ACKNOWLEDGMENTS

This work was partially funded by the grant MTM2012-31118 from the Ministry of Economy and Competitiveness of Spain.

REFERENCES

Bender, R., and Grouven, U. (1998). "Using Binary Logistic Regression Models for Ordinal Data With Non-proportional Odds". *J Clin Epidemiol,* 51, 809-816.

Bercedis, P., and Harrell, F. (1990). "Partial Proportional Odds Models for Ordinal Response Variables". *Applied Statistics,* 39, 205-217.

Lando, D., and Skødeberg, T.M. (2002). "Analyzing Rating Transitions and Rating Drift with Continuous Observations". *Journal of Banking and Finance,* 26, 423-444.

McCullagh, P. (1980). "Regression Models for Ordinal Data (with Discussion)". *Journal of the Royal Statistical Society, Series B*, 42(2), 109-142.

McNeil, A.J., and Wendin, J.P. (2006). "Dependent Credit Migrations". *Journal of Credit Risk*, 2(3), 87-114.

Wozabal, D., and Hochreiter, R. (2012). "A Coupled Markov Chain Approach to Credit Risk Modeling". *Journal of Economic Dynamics and Control*, 36, 403-415.

Yee, T.W. (2010). "The VGAM Package for Categorical Data Analysis". *Journal of Statistical Software*, 32, 1-34.

12. SOME PARAMETRIC TESTS IN PARTIALLY-NON-REGULAR LOG-EXPONENTIAL MODELS

I. Barranco-Chamorro
University of Sevilla, Department of Statistics and Operations Research (Faculty of Mathematics), Spain, chamorro@us.es

ABSTRACT

In this paper we propose asymptotic methods to carry out hypothesis tests on values of a parametric function when sampling from partially-non-regular log-exponential models. Specifically, we study the likelihood ratio, Wald, and score tests. Our results are applicable to the two-parameter exponential, power-function and Pareto distributions. Simulations and applications to lifetime data analysis and economics are included.

1. INTRODUCTION

Location-scale models are characterized by having probability density function (pdf), $f(\cdot)$, given by

$$f(y;\theta,\sigma) = \frac{1}{\sigma} f_0\left(\frac{y-\theta}{\sigma}\right), \qquad y \in \mathbb{R} \tag{1}$$

where $\theta \in \mathbb{R}$ is a location parameter, $\sigma > 0$ 0 is a scale parameter, and $f_0(\cdot)$ is a fully specified pdf. If > 0 is a random variable (rv) and $Y = \ln X$ has pdf of the form (1) then X has a *log-location-scale distribution*. The Weibull, log-logistic, and log-normal are all log-location-scale distributions; the corresponding location-scale distributions are the extreme value, logistic and normal, respectively. These models can be considered *regular* for inferential purposes and have been extensively studied, see for instance Lawless (2003) and references therein.

In this paper, we focus on a subclass of log-location-scale models that can be called *partially non-regular*. The point is that, unlike the regular models previously cited, the support of distributions we consider *depends* on the unknown location parameter. So we will have that the estimator of the location parameter is non-regular, whereas the estimator of the scale parameter behaves in a regular way. In this context, the term *partially non-regular* was proposed by Dubinin and Vardeman (2003). There exist hardly papers by using this approach to deal with inferential issues.

We highlight Smith (1985), Dubinin and Vardeman (2003) and BarrancoChamorro and Jiménez-Gamero (2012). Obviously, the mixture of estimators we have in these models leads to complications with respect to inferential purposes. The term *log-exponential distributions* was proposed in Barranco-Chamorro and Jiménez-Gamero (2012) for the subclass of loglocation-scale models that can be related, via the logarithm transformation, to the two-parameter exponential distribution. As important cases of them, we can cite the power-function (PF) and the classical Pareto models with unknown scale and shape parameters and whose support depend on the scale parameter.

The problem we face in this paper is the following one. We consider a simple random sample (srs) of size n from (1) in the partially non-regular case. Let us denote by $\hat{\theta}_n$ and $\hat{\sigma}_n$ the MLEs of the unknown parameters, and consider h a real-valued smooth function of $h(\hat{\theta}_n, \hat{\sigma}_n)$. Our aim is to propose several asymptotic procedures to test H_0: $h(\theta, \sigma) = 0$ versus H_1 : not H_0. Specifically, we consider the likelihood ratio, Wald, and score tests. Results are obtained when sampling from a two-parameter exponential distribution. Later, these procedures are extended to models that can be related via a logarithm transformation to the two-parameter exponential distribution. Finally, the performance of the three methods is compared through examples and simulations carried out by using R. Practical applications to reliability, lifetime models, and economics are included.

As precedent of this paper, we can cite Barranco-Chamorro and Jiménez Gamero (2012), where we gave conditions on the parametric function h to get approximations to the moments of $h(\hat{\theta}_n, \hat{\sigma}_n)$, and got the limiting distribution of $h(\hat{\theta}_n, \hat{\sigma}_n)$. It was proven that, in general, the asymptotic behaviour of $h(\hat{\theta}_n, \hat{\sigma}_n)$ is dominated by the behaviour of the 'regular' estimator $\hat{\sigma}_n$. However there exist other possible situations in

which the asymptotic properties of the non-regular estimator should be taken into account. In this paper, we propose and study other asymptotic approaches to deal with inferential issues in these models. We highlight that our approach is based on the knowledge of the limiting joint distribution of $h(\hat{\theta}_n, \hat{\sigma}_n)$, that in this case is no longer bivariate normal.

2. PARTIALLY NON-REGULAR LOG-EXPONENTIAL MODELS

In this section we introduce some notation and basic results in models we are dealing with.

Definition 2.1. A rv Y follows a two-parameter exponential distribution, $E(\theta, \sigma)$, if its pdf is given by

$$f(y) = \frac{1}{\sigma} e^{-(y-\theta)/\sigma} I_{[\theta,+\infty)}(y), \quad \theta \in \mathbb{R}, \quad \sigma \in \mathbb{R}^+, \tag{2}$$

where θ is a location parameter that also determines the support of the distribution, σ is a scale parameter, and $I_{[\theta,+\infty)}(\cdot)$ denotes the indicator function of $[\theta, +\infty)$.

Given $Y_1, \ldots, Y_n$, a simple random sample of size n from (2), the MLEs of θ and σ are $\hat{\theta}_n = Y_{(1)} = \min_{1 \le i \le n}\{Y_i\}$, respectively. $\hat{\theta}_n$ is non-regular for inferential purposes whereas $\hat{\sigma}_n$ is regular.

Definition 2.2. Partially-non-regular log-exponential models are models such that the logarithm transformation of original variables converts the order statistics from these distributions into the order statistics from an $E(\theta, \sigma)$ distribution.

Examples of them are the Power-function and Pareto distributions defined by the following pdf's.

- Power-function, $PF(\xi, \alpha)$, with $f(x) = \alpha \frac{x^{\alpha-1}}{\xi^\alpha} I_{(0,\xi]}(x), \quad \xi \in \mathbb{R}^+, \quad \alpha \in \mathbb{R}^+$.
- Pareto, $Pa(\xi, \alpha)$, with $f(x) = \alpha \frac{\xi^\alpha}{x^{\alpha+1}} I_{[\xi,\infty)}(x), \quad \xi \in \mathbb{R}^+, \quad \alpha \in \mathbb{R}^+$.

In both cases α is a shape parameter, and ξ is a scale parameter that also determines the support of the distribution. When sampling from these models, results obtained for an exponential distribution can be applied with minor modifications. We must apply the logarithm transformation to the variables and reparameterize the original parametric function in terms of $\sigma = \alpha^{-1}$. Details can be seen in Barranco-Chamorro and Jiménez-Gamero (2012).

3. TEST PROCEDURES

In this section we present the three test procedures we are going to study in detail. First, we develop results for an $E(\theta, \sigma)$ distribution. They can be extended to log-exponential models.

Let us consider a hypothesis test of the form

$$\begin{aligned} H_0 &: h(\theta, \sigma) = 0 \\ H_1 &: h(\theta, \sigma) \neq 0 \end{aligned}$$

where (θ, σ) are unknown parameters of an $E(\theta, \sigma)$ distribution, and h is a real-valued smooth function.

3.1. Wald Method

This method is based on the following result.

Theorem 3.1. Let $Y1, \ldots, Yn$ be a srs from the $E(\theta, \sigma)$ distribution defined in (2), and h a parametric function admitting a first order Taylor expansion at $(\theta, \sigma) \in \Theta$. If $D_2h(\theta, \sigma) \neq 0$ at (θ, σ) then

$$\frac{\sqrt{n}}{\hat{\tau}_n} \left\{ h(\hat{\theta}_n, \hat{\sigma}_n) - h(\theta, \sigma) \right\} \xrightarrow{\mathcal{L}} Z, \tag{3}$$

where $Z \sim N(0, 1)$, $D_2h(\theta, \sigma)$ denotes the first partial derivative of h with respect to σ evaluated at a given $(\theta, \sigma) \in \Theta$, i.e. $D_2h(\theta, \sigma) = (\partial/\partial\sigma)h(\theta, \sigma)$, and $\hat{\tau}_n$ is a consistent

estimator of $\sigma D_2 h(\theta, \sigma)$. That is, if we denote by $\hat{h}_n = h(\hat{\theta}_n, \hat{\sigma}_n)$ then the Wald test statistic is

$$|Z_W| = \frac{\left|\hat{h}_n - h\right|}{\widehat{se}(\hat{h}_n)} \tag{4}$$

where $\widehat{se}(\hat{h}_n)$ is the estimated standard error of $\hat{h}_n$. Under H_0, Z_W follows a $N(0, 1)$ distribution. The test is carried out by comparing the ZW -value obtained from (4) to the appropriate value from the $N(0, 1)$.

We highlight that (4) is easy to apply, but it presents some deficiencies.

3.2. Score Method

This method is based on the calculus of the MLE for the constraint model proposed in H_0. This MLE is denoted by $(\hat{\theta}_h, \hat{\sigma}_h)$. Analogously to the previous result, we have a standard normal distribution for the test statistic based on $(\hat{\theta}_h, \hat{\sigma}_h)$.

3.3. Likelihood Ratio Test (LRT)

The LRT statistic is given by

$$\Lambda_h = \frac{\sup\limits_{(\theta,\sigma)} \prod\limits_{i=1}^{n} f_{(\theta,\sigma)}(X_i)}{\sup\limits_{(\theta,\sigma) \text{with } h(\theta,\sigma)=0} \prod\limits_{i=1}^{n} f_{(\theta,\sigma)}(X_i)}.$$

In order to apply this method, we need to obtain the MLEs of (θ, σ), $(\hat{\theta}_h, \hat{\sigma}_h)$, and the MLEs of (θ, σ) under the constraint proposed in H0, $(\hat{\theta}_h, \hat{\sigma}_h)$. This test is usually given in terms o f $\lambda_h = 2 \ln \Lambda_h$, whose limiting distribution under H_0 is

$$\lambda_h \xrightarrow{\mathcal{L}} \chi^2_k, \qquad k = 1, 2, \tag{5}$$

that is, a chi-square distribution with k degrees of freedom, where k depends on the gradient vector of h, ∇h, at (θ_0, σ_0), with (θ_0, σ_0) a vector of parameters satisfying the constraint proposed in $H0$, $h(\theta_0, \sigma_0) = 0$.

The results obtained for the $E(\theta, \sigma)$ distribution are extended with minor modifications to log-exponential models. To conclude, we highlight that the problems we have faced in this section are not obvious. In Section 4, we include some applications and simulations that illustrate the performance of all these methods.

4. APPLICATIONS

In this section, we point out that the models considered in this paper have been extensively applied in reliability, lifetime data analysis, and economics. For this reason, we consider parametric functions of interest in these fields, such as quantiles, and functions that allow us to estimate the Lorenz curve and the Gini index. By using R, simulation studies have been carried out. They allow us to asses and compare the performance of the methods proposed. We conclude that, even for moderate sample sizes, our results perform quite well.

5. CONCLUSIONS

It can be said that the essential contribution of this paper is to show how to carry out asymptotic tests when we have regular and non-regular estimators. This is a complicated parametric problem that, in this way, it can be handled satisfactorily. Quite often the sample size is large enough to apply these results. We have also highlighted the practical importance of distributions we are dealing with. We can find applications of them to diverse fields of lifetime data analysis and economics. New uses of these models are continuously emerging in practice. For instance, Pareto distribution has recently been proposed to describe the traffic of files in the Internet, see for instance (Chabchoub *et al.*, 2010). Our results can be applied to deal with problems emerging in these new settings. We have shown, through simulations, that our results perform quite well, and may be compared to other popular techniques used in practice. The method is important and other inferential problems could be tackled similarly.

ACKNOWLEDGMENTS

This research has been partially supported by grant UJA2013/08/01.

REFERENCES

Akahira, M., and Takeuchi, K. (1995). *Non-Regular Statistical Estimation.* Lecture Notes in Statistics 107. Springer: New York.

Arnold, B.C. (1983). *Pareto Distributions.* International Cooperative Published House: Fairland, MD.

Barranco-Chamorro, I., and Jiménez-Gamero, M.D. (2012). "Asymptotic Results in Partially Non-Regular Log-Exponential Distributions". *Journal of Statistical Computation and Simulation,* 82, No. 3, 445-461.

Chabchoub, Y. Fricker, C., Guillemin, F., and Robert, P. (2010). "On the statistical Characterization of Flows in Internet Traffic with Application to Sampling". *Computer Commun.* 33, 103-112.

Dubinin, T.M., and Vardeman, S.B. (2003). "Likelihood-Based Inference in Some Continuous Exponential Families with Unknown Threshold Parameters". *JASA,* 98, No. 463, 741-749.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data (2nd edition).* Wiley: Hoboken.

Smith, R.L. (1985). "Maximum likelihood Estimation in a Class of Non-Regular Cases". *Biometrika,* 72(1), 67-90.

13. RISK IN ECO-INNOVATION INTRODUCTION AT ENTERPRISES

P. Bartoszczuk
Warsaw School of Economics, Poland
pbarto1@sgh.waw.pl

ABSTRACT

The goal of this paper is to present eco-innovation implementation risk at enterprises. Eco-innovation is rather modern term and can be a method to solve emerging environmental problems as consequences of economic growth. As for any innovation, eco-innovations have several types and can result therefore in a new or significantly improved product (good or service), process, a new marketing or organisational methods. Ecoinnovation should be seen as an integral part of innovation efforts across all the economy sectors. European countries observe many barriers for implementation of eco-innnovation, mainly associated with the high investments risk and limited interest.

1. INTRODUCTION

The European Union goal is to have an economic growth with ecoinnovations (EIO, ETAP). However, there are significant barriers to implement eco-innovations that need to be overcome before they are embraced by consumers. These are mainly caused by high risk of market and uncertain return on investment. Some environmental innovations require a lengthy period of time before they are adopted, which are directly related to their diffusion rate and diffusion path. The "New Member" countries in European Union are on a path to introduce broadly environmental technologies. We argue that eco-innovation in countries' companies is a key to the transformation towards green economy. Both current and expected regulations are particularly important for pushing firms to reduce air and water pollution, decrease

hazardous substance risk and increase recyclability of products. Cost savings are an important incentive for reducing energy and material use. Customer requirements are very important source for eco-innovations, particularly with regard to products with improved environmental performance and process innovations that increase material efficiency, reduce energy consumption and waste disposal.

This paper consist of six chapters including introduction and conclusion. First, I review literature about Eco-innovation and analyse how different stakeholder can contribute to building a green economy through eco-innovation. In the second chapter I analyze eco-innovation definition, and in the third pull factors. The fourth chapter deals with eco-innovation gap and in five I give some practical examples. In conlusion I underline incentives and barriers of ecoinnovation development in European companies.

2. WHAT IS ECO-INNOVATION?

First, we have to underline that no agreed definition of ecoinnovation has been developed until today. First definition of eco-innovation was formulated by C. Fussler i T. James (1996, p. 364), who described them as bringing profit both to producers and consumers, and simultaneously decreasing negative impact on environment. Since the 90's many definitions were generated but this general category is very problematic to define, requiring both theoretical knowledge and practical research. In the last years this concept is becoming very attractive among researchers, entrepreneurships, policy makers. Many previous work on eco-innovation have focused only on environmentally motivated innovation, neglecting completely the environmental gains from "normal" innovations (Kemp, 2008). Here, normal innovation is described as innovation for normal market reasons of saving costs or providing better services to users.

Instead of viewing the environment as just a source of raw materials or as an external challenge, companies in the future will internalise environmental sustainability in how they meet customer needs. Businesses will change the rules of the game by changing how they create, deliver eco-innovative product and capture value.

The outcomes of the ecoinnovation are very important to the decision makers. Traditional environmental policies and measures focus on dealing with specific problems. For instance, it has cleaned up water pollution, recycled certain products, contributed to climatic change decreasing risk. However, since traditional problem solving begins after recognising a problem's existence, risk, such policies are neither helpful on a systems level, nor are they preventive in a general. Solving individual problems only by specialised environmental technologies can even lead to risk of other problems, in particular those as yet undiscovered. Today, more than 95% of the resources lifted from nature are wasted before the finished goods reach the market. And many industrial products such as cars demand additional natural resources while being used. Sufficiently decoupling production and consumption from nature requires new philosophy, systems, goods, services, processes, and procedures for meeting human needs.

Besides the environmental gain, there should be a novelty, either from the development point of view or an adoption point of view. According to the Oslo Manual (OECD, 2005), innovation is defined as the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practice. What is important is that OECD makes a distinction between innovations new to the world and those new to the adopter.

Although quite many companies implement eco-innovation, but the majority either still do not eco-innovate due to risk or the material savings gained due to innovation are not important. Nevertheless there exists solid, untapped, often unrealized potential for eco-innovation in the EU. Around one fourth of innovating companies in the EU-27 have reported introducing eco-innovation to reduce material use in years 2006-2008, according to the Community Innovation Survey/CIS/ (Eurostat, 2008).

3. ECO-INNOVATION MAIN PULL FACTORS

Kempton *et al.* (1991), and Jaffe and Stavins (1994); Kenzig and Wustenhagen (2008) stated that there are some characteristics of eco-innovations influence market pull factors: "Costumer investment decisions regarding eco-innovations are characterized by":

- Different investment profiles over time (higher purchase price and set-up costs) and lower costs which means a different return on investment profile.

- Information asymmetries due to search experience, and credence attributes. Eco-innovations are hampered by the risk stigma. This is due to information asymmetries since the necessary information is needed to overcome this prejudice/ this point of view.

- Externalities (e.g. environmentally sound alternatives imply a higher collective benefit but lower or equal private benefits than conventional alternatives) Eco-innovations may have the characteristic that the party taking the investment is not the party that captures/ receives the correlated benefits.

- Infrequent decisions, meaning an extensive decision-making process which implies high involvement, high cognitive effort, high risk, and a substantial need for information due to limited experience. Customers interested in eco-innovations are usually not yet experienced in the decision making process since eco-innovations, most of the time, consist of a relatively new, high risk and undiscovered market or technique. Dutch experts of commercial parties (business organisations, scientists and advisors) do not really see any major differences between eco-innovations and regular innovation, with the exception of systematic innovations (Bertens Statema, 2011).

4. THE GAP IN ECO-INNOVATION

There is a gap both in the scale of eco-innovative activities, with large differences between countries, sectors, and sizes of companies as well as a gap in the scope of eco-innovation changes, with a tendency towards more incremental rather than radical changes introduced to the market.

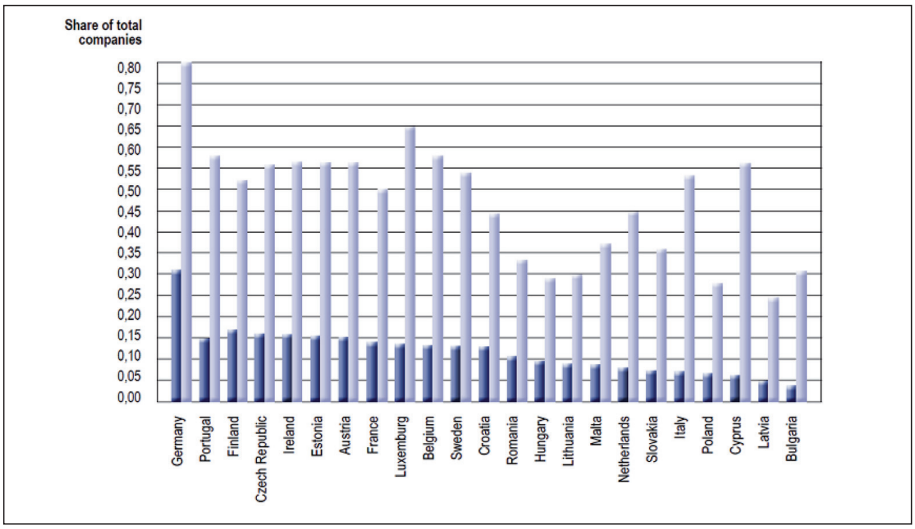


Figure 1. Share of innovating companies reducing material use per unit of output, 2006-2008. Source: EIO (2012), p. 7.

The gap between innovators and eco-innovators is significant across the EU. In Germany around 39% of innovators are eco-innovating, but only 10% in Cyprus and Bulgaria (see Figure 1 for further differences). If one looked at the share of eco-innovators in total companies, only around 15% of companies across the EU have reported eco-innovative activity focused on material efficiency. In Poland there is little innovation (located on the penultimate place in the ranking). Eco-innovators are about 22% of the total innovators in Poland.

Poland is among the countries scoring persistently low in the European Eco-innovation Scoreboard since 2010 (EIO, 2012). In the 2013 edition Poland came second last among the EU countries with the score below the EU average (45 out of 100). The overall low score in the index, especially in terms of eco-innovation inputs and outputs, reflects the low innovativeness of the country in general (Szpindor and Sniegocki, 2012).

The low score in the eco-innovation activities may be underestimated. According to a recent Eurobarometer survey More than 9 out of 10 Polish companies surveyed by the Eurobarometer in 2011 claimed to have introduced at least one eco-innovation

in the past two years (Gallup, 2011). This seems rather plausible, as the basic improvements in energy efficiency were part of the overall modernisation in the Polish economy in the 1990. Companies tend to consider these changes 'incremental rationalisation' rather than an eco-innovation. While we may expect improvements in eco-innovation inputs and activities in the coming years, the improvement of resource efficiency outcomes would require lasting systemic innovation efforts before any substantial improvements become visible in statistics. The relatively poor performance is caused by many structural factors rarely associated with the natural environment (e.g. science-industry collaboration, absorption capacity of SMEs, the lack of economic and fiscal incentives, high investment risk, etc.). These are sometimes referred to as 'bottlenecks' in the innovation system. Removing systemic barriers to innovation while adding a strong sustainable strategic direction in the policy framework could reinvigorate the ongoing process of socioeconomic transformation in Poland.

5. EXAMPLES OF THE ECO-INNOVATION

The country has outstanding examples of eco-innovations in energy and water management, hazardous waste treatment, solar energy, green banking and coke industry and a number of eco-innovation related programmes and initiatives within clusters. On the other hand, ecoinnovation does not constitute a driving force for new business opportunities in Poland. One of the examples of successful ecoinnovations is a device for water aeration. POLIMAT EKO designed original solutions for a tool used to aerate and de-gas waters in reservoirs (lakes, ponds, rivers) operating as a sprinkler and aerator for water. In the water treatment technology, aerators are used for de-ironing and demanganation of water. Application of aerators decrease the volume of pollution dissolved in water.

Aerating devices of Polimat Eko are applied in water (rivers, lakes, ponds, water holes, reservoirs), in sewage treatment plants and in industry. These are standing or floating aerators with relevant capacities. They are made of materials which do not require maintenance; and resistant to corrosion and subject to recycling and powered with electric engines of small power; the aeration process is fully automated and may be optionally monitored by the company. With relatively small sizes,

these devices are very efficient in the aeration process. The artificial aeration system proposed by POLIMAT EKO does not cause changes in water temperature and in consequence leads to reduction in internal enrichment of water with bio-genes from bottom sediment. Aeration devices of Polimat Eko participate in an innovative reclamation project of Lake Parnowo.

6. CONCLUSION

The main barrier to eco-innovation in Central Europe, including Poland is the overall low level of innovativeness of the Polish economy (Szpindor and Sniegocki, 2012). The most significant barriers faced by companies that implement eco-innovation were lack of funds, difficult access to capital, the relatively high cost of eco-innovative technologies, risk of market demand and uncertain return on investment, the lack of economic and fiscal incentives, administrative barriers as being a problem, often in relation to the risk-averse public procurement practices. The lack of sufficient knowledge may lead to the perceptions of eco-innovation as a cost rather than potential benefit to the company, which removes the main incentive to invest in innovation activity (Szpindor and Sniegocki, 2012). Another barrier signalled by companies developing environmental technologies was the risk that their customers did not consider environmental benefits when making purchasing decisions. Unexpectedly, the great majority of the same sample would routinely point to the environmental benefits of their technologies in their marketing activities.

REFERENCES

Bertens, C., and Statema, H. (2011). *Business Models of Eco-Innovations. An Explorative Study into the Value Network of the Business Models of Eco-Innovations and some Dutch Case Studies.* Zoetermeer.

EIO (2012). *The Eco-Innovation Gap: An Economic Opportunity for Business.* Eco-Innovation Observatory. Funded by the European Commission, DG Environment, Brussels.

Gallup (2011). *Attitudes of European Entrepreneurs Towards Ecoinnovation.* Analytical report, Flash Eurobarometer 315.

Jaffe, A.B., Newell R.G., and Stavins, R.N. (2002). "Environmental Policy and Technological Change", *Environmental and Resource Economics,* 22,1-2, 41-70.

Kemp, R., and Pearson, P. (2008). *Final Report MEI Project about Measuring Eco-Innovation,* Maastricht, www.merit.unu.edu\MEI .

Kaenzig, J., and Wüstenhagen, R. (2010). "The Effect of Life Cycle Cost Information on Consumer Investment Decisions Regarding EcoInnovation". *Journal of Industrial Ecology,* 14, (1), pp. 121-136.

Kempton, W. (1991). "Lay Perspectives on Global Climate Change". In Edward L. Vine, Drury Crawley and Paul Centolella (eds.), *Energy Efficiency and the Environment: Forging the Link,* 29-69. Washington, D.C.: American Council for an Energy Efficient Economy.

ETAP (2010). *Clean, Clever, Competitive Eco-Innovation for a Sustainable Future,* newsletter issue 18, European Commision, Brussels, Belgium.

Oslo manual. The Measurement of Scientific and Technological Activities Proposed Guidelines for Collecting and Interpreting Technological Innovation Data (2005). OECD, www.oecd.org

Szpindor, A., and Śniegocki, A. (2012). *Ekoinnowacje w Polsce. Stan obecny, bariery rozwoju, możliwosci wsparcia.* Instytut Badan Strukturalnych.

14. RISK BEHAVIOR BEHIND DISTORTION RISK MEASURES

J. Belles-Sampera
Department of Econometrics, Riskcenter-IREA,
University of Barcelona, Spain,
belles.sampera@ub.edu

ABSTRACT

This article characterizes the risk attitude involved in distortion risk measures. First, it is shown that the area of the distortion function provides valuable information related to the implicit risk attitude in aggregate terms, which leads to the definition of aggregate risk attitude. In the second stage a graphical analysis based on the distortion function is carried out to provide a local description of the underlying risk behavior. Here, the concepts of absolute risk attitude and relative risk attitude arise in a natural manner.

1. INTRODUCTION

This article pursues to contribute into the study of the risk behavior in risk assessment. Properties and limitations of risk measures have been extensively studied in the literature. Up to our knowledge, however, scarce attention has been paid to the risk attitude that the decision maker is assuming when a risk measure is chosen. This study analyses the risk behavior that is implicit in the agent applying a particular risk measure. The attention is focused on the family of distortion risk measures introduced by Wang (1995, 1996). The characterization of the implicit risk attitude that is involved in a distortion risk measure is carried out by means of the analysis based on the distortion function.

In particular, it is investigated the quantitative information related to the *aggregate risk behavior* linked to the risk measure provided by the area of the distortion function

linked to the distortion risk measure. The aggregate risk attitude only provides a partial portrait of the underlying risk attitude. It is reasonable to suppose that decision makers are not worried about all random events in the same way. Decision makers frequently give different treatments to different random events. Note that some of these events can represent benefits or affordable losses. Therefore, the *aggregate risk behavior* of a risk measure has to be completed with local information about the implicit risk attitude in front of a particular random event. A quotient function based on the distortion function associated to the risk measure is defined with this purpose. The quotient function is graphically analyzed to investigate the risk attitude of the agent at any point of the survival distribution function when using a certain risk measure. The graphical evaluation of the risk-appetite pattern of the manager in the range of feasible values is the basis of the definition of two concepts, absolute risk behavior and relative risk behavior.

2. DISTORTION RISK MEASURES

A distortion function is a non-decreasing and injective function g from [0, 1] to [0, 1] such that $g(0) = 0$ and $g(1) = 1$. Consider a probability space and the set of all random variables defined on this space. Given one of these random variables X, the value $\rho g(X)$ that a distortion risk measure[2] returns when applied to X may be understood as the value of the asymmetric Choquet integral of X with respect to a capacity, a capacity which is build by distorting the survival probability of X with the distortion function g, i.e.

$$\rho_g(X) = \int X d(g \circ P).$$

Most frequently used distortion risk measures are the quantile-based risk measures Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR). Let consider $\mathrm{VaR}\alpha$ and $\mathrm{TVaR}\alpha$ risk measures which are defined for a random variable X as $\mathrm{VaR}_\alpha(X) = \inf\{x \mid F_X(x) \geq \alpha\} = F_X^{-1}(\alpha)$ and $\mathrm{VaR}\alpha\ (\mathrm{X}) = \dfrac{1}{1-\alpha}\displaystyle\int_\alpha^1 VaR_\lambda(X)\, d\lambda$, respectively, where $0 \leq \alpha \leq 1$ is a confidence level.

[1] As introduced by Wang (1995, 1996).

The relationship between the VaR and TVaR risk measures and discrete Choquet integrals has been shown in Belles-Sampera *et al.* (2013). Both $\mathrm{VaR}\alpha\,(\mathrm{X})$ and $\mathrm{TVaR}\alpha\,(\mathrm{X})$ may be understood as Choquet integrals with respect to capacities $\nu = \psi\alpha \circ P$ and $\tau = \gamma\alpha \circ \mathrm{P}$, respectively, where P is the $\psi_\alpha(u) = \mathbb{1}_{[1-\alpha,1]}(u)$ and $\gamma_\alpha(u) = \frac{u}{1-\alpha} \cdot \mathbb{1}_{[0,1-\alpha)}(u) + \mathbb{1}_{[1-\alpha,1]}(u)$ distortion functions associated to these risk measures.

Many articles have appeared in recent years that pay attention to risk measures based on distortion functions or on generalizations of the quantiles. Some examples of the first group are Zhu and Li (2012) and Belles-Sampera *et al.* (2014). Bellini and Gianin (2012) and Bellini *et al.* (2014) fit to second group. An interplay between both groups is found in Dhaene *et al.* (2012) and Goovaerts *et al.* (2012).

3. RISK ATTITUDE

It is analyzed here the risk behavior of agents when they select a particular distortion risk measure to assess the risk. Note that the distortion risk measure is defined as $\rho\,(X) = C_g \circ P\,(X)$, where $C_g \circ P\,(X)$ is the Choquet integral with respect to capacity $g \circ P$ being P the cdf of X. So, the role played by the distortion function g is to distort the survival probability of X. The mathematical expectation of X can be understood as a particular case of distortion risk measure such that $\mathbb{E}\,(X) = \mathcal{C}_{id \circ P}\,(X)$, where the distortion function is the identity function id. Indeed, the value of a distortion risk measure $\rho\,(X)$ may be interpreted as the expectation of X given that the survival probability of X has been previously distorted by the function *g*. So, the area under *g* can be understood as an indicator of the aggregated risk behavior of an agent.

Decision-makers are frequently classified as either risk averse, risk neutral or risk loving agents. Note that a risk neutral agent would not distort the survival distribution function, so the associated distortion function linked to aggregated risk neutrality should be the *id* function. The area under the *id* function is one half and this value could be used as a benchmark of aggregated risk attitude or behavior. An aggregated risk averse agent would make an upper distortion of the survival distribution in accumulated terms. In consequence, the area under *g* of an aggregated risk averse agent should be larger than one half. Similarly, an agent would have aggregated risk appetite if the area under *g* is lower than one half.

A graphical analysis is proposed to complement the examination of risk attitudes of decision makers when using distortion risk measures. An indicator of the area under the distortion function evaluates the accumulated distortion performed over the survival distribution function. Unfortunately, this measure does not have into account which part of the survival distribution function was distorted. From the point of view of the manager, of course, it is not equivalent to distort the survival probability associated to the right tail of the random variable linked to losses or to distort the left tail. Additionally, all distortion functions with an area equal to one half would be linked to aggregated risk neutrality, where the id function is only a particular case.

An option is to define the risk behavior in absolute terms. An absolute *risk neutral agent* should be the decision-maker who does not distort any survival probability and, then, who use the id function as associated distortion function, i.e. $g(u) = id(u) = u$ for all $0 \leq u \leq 1$. An absolute risk averse agent should have associated a distortion function g such that $g(u) > u$, for all $0 \leq u \leq 1$. And, in the same line, an absolute risk loving agent should have a distortion function g such that $g(u) < u$, for all $0 \leq u \leq 1$. This definition of risk behavior is in absolute terms in the sense that the relationship of ordering between $g(u)$ and u must be fulfilled in the whole range [0, 1]. Note that these considerations lead to a more restrictive definition of risk behavior than the previous *aggregated risk behavior*. Under the definition of aggregated risk behavior it is only required to fulfill the classification criterion in aggregated terms. So, all absolute risk averse agents are aggregated risk averse agents but the opposite does not hold.

The absolute risk behavior definition implies that the implicit risk attitude of an agent is invariable over the range of values. Nevertheless, there are no reasons to have an unique risk attitude on the whole range. An agent could have a different risk-behavior depending on the interval of values under consideration. The risk attitude implicit in frequently used risk measures is not invariant. It is the case, for instance, of VaRα. When using VaRα, a risk averse behavior is involved in the interval $[1 - \alpha, 1)$, but a risk loving attitude is associated to the interval $(0, 1 - \alpha)$. So, an absolute risk behavior can not be linked to the VaRα risk measure. In those situations that the implicit risk attitude is not constant, the absolute risk behavior criterion would not be an adequate classification.

The study of the risk behavior of agents when used a particular risk measure is often more explanatory when is locally investigated in each particular area. Let us define the function *Qg (u)* on $u \in [0, 1]$ as the quotient between the distortion function g and the identity function, $Q_g(u) = \frac{g(u)}{u}$ for all $0 \leq u \leq 1$. The *Qg (u)* allows the analysis of the agent's risk behavior at any survival probability point. It takes non-negative values. This quotient function provides a function on survival probabilities u which describes the distortion factor applied at each u level by g. The quotient Qg is a quantifier of the relative risk behavior of the agent at any point. The quotient value represents the relative risk attitude of the decision maker in comparison to the risk neutral attitude of the agent who is confident with the survival probability. An agent is risk neutral, risk averse or risk loving at point u if *Qg (u)* is equal, higher or lower than one, respectively. The graphical analysis consists of plotting *Qg (u)* against the identity function, for all $0 \leq u \leq 1$. Two examples are shown in Figure 1, where quotient functions associated to VaRα and to TVaRα are displayed.

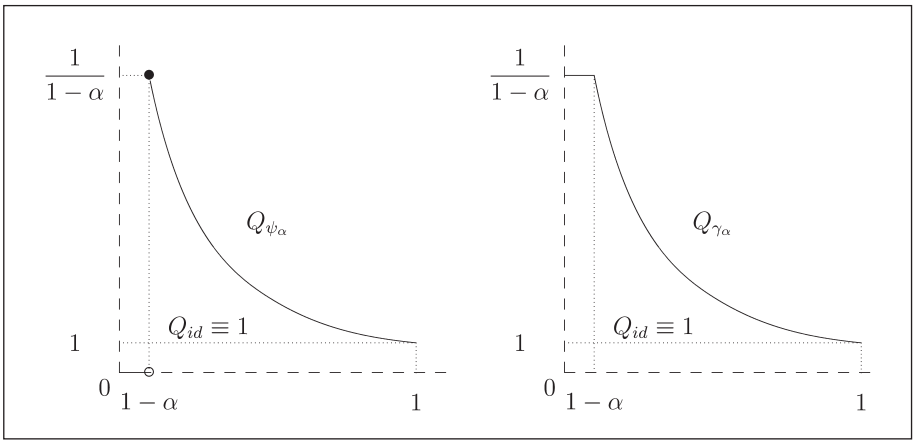


Figure 1. Quotient functions for VaRα (left) and TVaRα (right). Note that they only differ on the interval [0, 1 – α). Q_{id} is the quotient function of the mathematical expectation $\mathbb{E}$.

Let emphasize that the quotient $1/u$, for all $0 \leq u \leq 1$, marks out the maximum risk aversion frontier (as it can be deduced from Figure 1). The function *Qg (u)* computes the ratio between the distorted survival probability and the survival probability, so $1/u$ is the maximum value attainable by *Qg (u)*. Note that the function *Qg (u)* is equal to 1 in the whole range for a completely risk neutral agent in absolute terms *(absolute*

risk neutral agent }. So, values $0 \leq Qg\,(u) < 1$ indicate a risk loving behavior of the agent at point u.

If we focus the attention on the $Qg\,(u)$ associate to the VaRα, it is shown that a radical risk attitude is implicit in the interval $[1-\alpha, 1)$ which varies to the opposite extreme position in the interval $(0, 1-\alpha)$. Indeed, maximum risk aversion is involved in $[1-\alpha, 1)$ and maximum risk loving attitude in $(0, 1-\alpha)$. Some similarities are found when $Qg\,(u)$ associated to the the TVaRα is examined. Two ranges involving a different risk attitude are distinguished as well. Maximum risk aversion is involved in the interval $[1-\alpha, 1)$ and a constant (non-boundary) risk aversion attitude is involved in $(0, 1-\alpha)$. Unlike the VaRα, an absolute risk averse behavior is associated to the TVaRα since $Qg\,(u) > 1$ or, equivalently, $g(u) > u$, for all $0 \leq u \leq 1$.

To conclude, we show that the quotient $Q_g\,(u)$ can be used to characterize the relative risk behavior of an agent at any point. The value of $Q_g\,(u)$ at a particular point depends on the distortion function, but on the original survival function as well. That is, risk attitude in the quotient function is contemplated on the size of the performed distortion (denominator) but also on the position in which this distortion is performed (numerator). Note that the area under the quotient function $Q_g\,(u)$ provides similar information than the area under g, but now it is expressed on the basis of risk neutrality. Indeed, the area under the quotient function can be interpreted as the area under a weighted distortion function, where weights $(1/u)$ are given to distorted values $Q_g\,(u) = g\,(u) \cdot 1/u$. Following this interpretation, higher weights are assigned to distortion function values associated to lower survival values. Therefore, the evaluation of the area under $Q_g\,(u)$ may be useful to analyze the aggregate risk behavior in those situations in which values of the distortion function are requested to be weighted indicating that risk aversion is negatively associated to the size of survival values. In this manner, the area under the quotient function is then interpreted as a weighted quantifier of the aggregated risk behavior, where an area equal to one indicates aggregate risk neutrality, an area larger than one aggregate risk aversion and an area lower than one aggregated risk appetite.

4. DISCUSSION

Distortion risk measures are widely used to assess risk in many contexts. The risk value returned by a distortion risk measure basically depends on the characteristics of the random variable, which are captured by the survival distribution, and the associated distortion function. These two parts are disaggregated and our attention is focused on the analysis of the distortion function. The distortion function can be understood as a weighting function of the survival probabilities. Therefore, any risk attitude implicit in a distortion risk measure is to some extent attached into its distortion function.

The size of the area under the distortion function reveals the accumulated distortion made over the survival distribution. So, this indicator informs about the aggregated risk attitude when a particular distortion risk measure is applied. Risk attitude implicit in a risk measure is not only determined by the area size of the distortion function. As important as the size of the aggregated distortion is how this total distortion is distributed through the range of values. The risk information provided by the area under the distortion function be complemented by the quotient function, which is the ratio between the distortion function and the identity function. We demonstrate that a graphical analysis of this quotient function gives local information of the relative risk behavior associated to the risk measure at any point of the range of values. In addition, the area under the quotient function provides similar aggregate risk information than the area under the distortion function, so this quotient function could replace the distortion function in the analysis of aggregate risk behavior, particularly if distorted values are requested to be weighted.

REFERENCES

Belles-Sampera, J., Guillén, M., and Santolino, M. (2014). "Beyond Value-at-Risk: GlueVaR Distortion Risk Measures". *Risk Analysis,* 34(1): 121-134.

Belles-Sampera, J., Merigó, J. M., Guillén, M., and Santolino, M. (2013). "The connection Between Distortion Risk Measures and Ordered Weighted Averaging Operators". *Insurance: Mathematics and Economics,* 52(2): 411-420.

Bellini, F., and Gianin, E.R. (2012). "Haezendonck-Goovaerts Risk Measures and Orlicz Quantiles". *Insurance: Mathematics and Economics*, 51(1): 107-114.

Bellini, F., Klar, B., Mueller, A., and Gianin, E.R. (2014). "Generalized Quantiles as Risk Measures". *Insurance: Mathematics and Economics*, 54: 41-48.

Dhaene, J., Kukush, A., Linders, D., and Tang, Q. (2012). "Remarks on Quantiles and Distortion Risk Measures". *European Actuarial Journal*, 2(2): 319-328.

Goovaerts, M., Linders, D., Van Weert, K., and Tank, F. (2012). "On the Interplay Between Distortion, Mean Value and Haezendonck-Goovaerts Risk Measures". *Insurance: Mathematics and Economics*, 51(1): 10-18.

Wang, S.S. (1995). "Insurance Pricing and Increased Limits Ratemaking by Proportional Hazard Transforms". *Insurance: Mathematics and Economics*, 17(1): 43-54.

Wang, S.S. (1996). "Premium Calculation by Transforming The Layer Premium Density". *ASTIN Bulletin*, 26(1): 71-92.

Zhu, L., and Li, H. (2012). "Tail Distortion Risk and its Asymptotic Analysis". *Insurance: Mathematics and Economics*, 51(1): 115-121.

15. ANALYZING AND MANAGING POLITICAL RISK

O.J. Bergfjord
Department of Economics, Bergen University College, Norway, ojb@hib.no

ABSTRACT

Political risk is the risk related to political decisions. Such risk comes in many forms, and is likely to be important for many firms and individuals. Still, political risk is relatively little analyzed, possibly because it is difficult to quantify and model. In this paper, I briefly discuss different types of political risk, and outline how such risk could be measured, analyzed and coped with**.** A more formal analysis will be provided in an extended version of the paper.

1. INTRODUCTION

Based on the traditional Knightian definition (Knight, 1921) political risk might seem like a contradiction in terms. Knight advocated a distinction between risk, with known or estimable probabilities and state payoffs, and uncertainty, where probabilities (and often potential states and/or state payoffs) are difficult to estimate. Political decisions are, by their nature, difficult to estimate probabilities for, and "political uncertainty" would, based on this, be a more appropriate term. However, the Knightian definition has become less important, and "political risk" seems to be the most common term these days. I will hence use the term "political risk" for the rest of the paper.

Some further clarification of the concept could also be necessary. Although political decisions could impact different stakeholders in different ways, this paper will be limited to the *economic* impact of such risk. Also, whereas individuals too face political risk, for instance related to changes in the tax system or the pension system, this paper will primarily consider the risk faced by for-profit firms.

Political risk is obviously important for many, if not most firms. Many examples are intuitively easy to grasp. All firms are affected by changes to the tax system. Many firms sell goods or services to the government, and particular industries has been substantially hindered (or supported) by government regulations or subsidies, for instance the tobacco industry or the renewable energy industry. An example of how such risk is perceived is provided by Bergfjord (2009), where fish-farmers were asked to rate various risk sources based on their perceived importance for them. Although fish-farming is not considered to be an industry particularly exposed to political decisions, the political risk factors were still considered the most important, ahead of various operational and financial risk factors.

Although intuitively important, surprisingly little research has been conducted on this particular type of risk. The most common perspective has been risk related to investment in foreign (developing) countries. As part of the finance literature, such studies often construct indices of risk levels for different countries, which in turn are used to calculate required rates of return for investments in different countries. Risk factors included in these analyses typically include risk of expropriation, corruption and various legal challenges (Clark, 1997).

Some work on political risk has also been done related to social security and pension benefits (see e.g. Shoven and Slavov, 2006), and within the political economy literature (see e.g. Persson and Tabellini, 2002 for an overview). However, most of this literature has different perspectives, for instance how political decisions can reduce or distribute risk.

The further structure of this short paper is as follows: First, I will discuss different types of political risk. Second, I will outline a framework for how such risk could be measured. Finally, I will discuss how economical agents could manage this risk, and whether/how governments should try to reduce the risk.

2. DIFFERENT TYPES OF POLITICAL RISK

It is difficult to present a complete list of political risk factors. The relevant risks and their relative importance will vary from firm to firm, over time, and between different countries.

It is nevertheless worth to try to sort risk factors along a few dimensions, not least because this makes it easier to discuss management strategies later. A first classification could be general risk vs specific risks. A hypothetical global value-added tax would be the ultimate general risk, as it would affect all firms, and it would affect them all in pretty much the same way. A government's expropriation of a firm's plant would be a very specific risk, as it would only affect this particular firm. Most risks are somewhere between these extremes. Some political decisions affect all or most firms in a country, and would thus be general on a national level. Other decisions affect one industry, whereas some decisions affect only a small number of firms within a industry.

A second classification could be based on the potential state payoffs. A potential minor change in the tax system would be a political risk, but would probably have a very small impact on most firms' profit. The potential expropriation of a plant would, on the other hand, represent a much larger risk for this particular firm, as the difference in payoffs between the two states "expropriation" and "not expropriation" would be much larger. It is here worth pointing out that typically, general risks (i.e., political decisions affecting a broad range of companies) are typically less dramatic for each firm.

Although not directly related to the two dimensions above, it is also worth contemplating whether risks are just or unjust. In Western democracies, most political decisions are legitimate, and the people's right to elect new governments which implement new policies is a valuable part of democracy. Hence, the risk of changes to the tax system, changes in regulations etc will generally be just, something firms will have to live with, and not necessarily something governments could and should try to eliminate. On the other hand, some risk factors, for instance corruption or illegitimate confiscation of property, would generally be considered unjust.

To summarize, risk factors could be viewed as general or more firm-specific; with small or large potential effects, and finally also as just or unjust.

3. MEASUREMENT OF POLITICAL RISK

In principle, political risk could often be measured like any other type of risk. Although I mainly have focused on downside risk so far, there is a potential upside to many political decisions too, which makes it possible, at least in principle, to measure political risk using standard subjective probabilities and utility calculations as described for instance by Gilboa (2011). Ellsberg's paradox (Ellsberg,1961) that uncertainty about probabilities is perceived to be an extra burden compared to taking on risks where the probabilities are known, will of course increase the perceived risk of many political decisions, but again, this is principally not different from many other risk factors.

There are, however, a few practical challenges which are particularly important when analyzing political risk factors. First, political decisions can be influenced. Both probabilities and state payoffs can be affected through for instance lobbying, yet the clear relationship between effort, probabilities and payoffs is often very difficult to establish. Second, most common techniques to estimate probabilities are harder for political decisions. Imagine, for instance, the risks related to a medical procedure[3]. Although the probability for success is based on many factors, and might appear difficult to estimate, large databases make it possible to build regression models which estimate probabilities based on most of the relevant variables (gender, age, medical history, surgeon performing the procedure etc). The problem with political decisions is first that it is very difficult to construct a database of relevant cases for comparison, and second, that if this is possible, all sorts of possible causal relations could usually be constructed between the different variables. Such "narratives" are indeed common to observe as the background for political decisions, but the interdependence of variables makes estimation of probabilities difficult.

[1] This example is used by Gilboa (2011).

Even with these practical challenges, there are many situations where "traditional" risk analysis could be used. Establishing estimates for probabilities remains the critical point. Where necessary, consultants or other external experts could be used to assist with this, but there are also many political decisions where probability estimates could be extracted from various markets. These include traditional financial markets, betting markets and prediction markets. The next step is to determine the state payoffs. This might be difficult if the state space is undetermined, but otherwise it should not be more difficult than for other risk factors. Finally, in order to quantify the effect of the risk factors, knowledge about risk aversion, for instance through certainty equivalents, is necessary. Again, this is difficult in practice, but no more so for political risk factors than for other risk factors.

4. MANAGEMENT FROM ECONOMIC AGENTS

One reason why political risk is considered important among firms is that such risk factors appear to be more difficult to handle than many other risk factors. The probability of many operational risks can be affected by operational procedures, and much risk can be eliminated through insurance. Many financial risks can also be eliminated through standard financial contracts, for instance futures or forward contracts. Political risks, on the other hand, are virtually impossible to insure against.

Thus, there remains four main alternatives on how to cope with such risk. The first alternative is the obvious - to do nothing at all. Although this might appear silly, there are some good reasons to consider this alternative, in particular in light of the theoretical finance literature. First, managing political risk has a clear cost in terms of labor and attention. One could often argue that this attention and labor is better spent elsewhere, for instance on working on profitable projects within the firm's regular business. Second, most firms have owners that equally well could manage this risk on their own. As an investor concerned with political risk, I could buy shares in one firm expected to benefit from a Republican president, and hedge this by also buying shares in a firm expected to benefit from a Democrat president. If one of these firms tries to eliminate their exposure to the election result, this will increase my overall risk as an owner rather than reducing it.

The second strategy could be termed protection. Through operational decisions, firms can in some cases reduce their exposure to political decisions. For instance, if a higher minimum wage is considered, firms could adjust by automating more of their production, thus reducing the impact of higher wages. If a firm perceives there is a risk a foreign government will expropriate their plant, it can adjust by reducing investment in this plant, and/or move all R&D-activities away, so the value of the plant as a stand-alone unit would be reduced.

The third alternative is trying to influence the decision, through various types of lobbying. This is of course common practice, and done both directly by individual firms, industry organizations and through professional lobbyists and agencies. Some potential pitfalls to this strategy are worth mentioning. First, it is often difficult to analyze the effects of lobbying. If a beneficial decision is made, it is often not clear whether this happened because of our lobbying, or whether the outcome would have been the same anyway. Second, lobbying is politically and culturally sensitive. If done wrong, it may backfire and create negative responses, both from the actual decision makers, but also among the general public.

The fourth option, less common in practice and often harder to implement, is hedging. If a firm is heavily affected by who is elected president, it could hedge some of this risk by buying contracts that pay out if the "wrong" candidate is elected. Although little used, such opportunities to use financial markets for hedging exist more often than one would first think, and could be worth exploring for firms wishing to reduce these risks.

5. GOVERNMENT ACTIONS

There are several good reasons why governments not necessarily should be concerned about reducing the political risk faced by firms. The first and foremost is that political decisions, including changes to previous decisions, are important for the democracy. If there were no risk of changes to the current situation, elections and democracies as we know it would serve no purpose. Second, some level of political risk could actually be benefical for society (Bergfjord and Brandt, 2009). The argument is that in some situations, the introduction of political risk will affect

agents' utility, such that different types of rent-seeking, which are negative for society, will no longer be profitable.

There are, however, many situations where governments should try to reduce the political risk. Although originally developed for a more specific problem, the fundamental insight from Kydland and Prescott (1977) is applicable to most such situations. "Rules rather than discretion" reduces the risk for most firms, and some practical applications are worth discussing. Using rules rather than discretion implies that political decisions are intended to be made for the long term, without sudden and frequent changes. In order to achieve this, it will often be useful to seek broad political compromises, unlikely to be overthrown by a new government. Furthermore, "rules rather than discretion" can be viewed as an argument in favor of broad, universal policies, rather than industry or company specific regulations. As mentioned above, this would typically reduce the political risk for any single firm, compared to a situation with lots of different, specific and everchanging regulations.

Finally, and obviously, the political risks described as unjust in the discussion above should be avoided. This would not only be good for firms avoiding unnecessary political risk, but also for the country's population as a whole.

REFERENCES

Bergfjord, O.J. (2009). "Risk Perception and Risk Management in Norwegian Aquaculture". *Journal of Risk Research,* 12(1), 91-104.

Bergfjord, O.J., and Brandt, U.S. (2009). "Regulations, Risk, and Rent Seeking Behavior". *Journal of Risk and Governance,* 1(3), 233-246.

Clark, E. (1997). "Valuing Political Risk". *Journal of International Money and Finance,* 16(3), 477-490.

Ellsberg, D. (1961). "Risk, Ambiguity, and the Savage Axioms". *Quarterly Journal of Economics,* 75, 643-669.

Gilboa, I. (2011). *Making Better Decisions.* Wiley-Blackwell, United Kingdom.

Knight, F.H. (1921). *Risk, Uncertainty and Profit.* University of Chicago Press, Chicago, USA.

Kydland, F.E., and Prescott, E.C. (1977). "Rules Rather than Discretion: The Inconsistency of Optimal Plans". *Journal of Political Economy,* 85, 473-491.

Persson, T., and Tabellini, G. (2002). *Political Economics: Explaining Economic Policy.* MIT Press, Massachusetts, USA.

Shoven, J.B., and Slavov, S.N. (2006). "Political Risk versus Market Risk in Social Security". *NBER Working Paper 12135.*

16. A BIVARIATE INAR(1) MODEL FOR INSURANCE CLAIM COUNTS

L. Bermúdez
Universitat de Barcelona, Riskcenter-IREA, Departament de Matemàtica Econòmica, Financera i Actuarial, Spain, lbermudez@ub.edu

M. Guillén
Universitat de Barcelona, Riskcenter-IREA, Departament d'Econometria, Estadística i Economia Espanyola, Spain, mguillen@ub.edu

D. Karlis
Athens University of Economics and Business, Department of Statistics, Greece, karlis@aueb.gr

ABSTRACT

A bivariate INAR(1) model is adapted to the ratemaking problem of pricing an automobile insurance contract with two types of coverages (third-party liability guarantee and other guarantees) taking into account both the correlation between claims from different type of coverage and the serial correlation between the observations of a same policyholder that are observed over time. A numerical application using an automobile insurance claims database is conducted and the main finding is that the improvement of the BINAR(1) model over the simplest models is very large implying that we need to consider both time and cross correlation to fit the data at hand.

1. INTRODUCTION

Automobile insurance market is a highly competitive market. Hence, modeling insurance claim count data is an essential part of tasks of actuaries.To ensure that

the exact weight of each risk is fairly distributed within the portfolio, a particularly complex pricing structure is designed by actuaries. Traditionally, ratemaking was separated into two steps called *a priori* ratemaking and *a posteriori* ratemaking. A thorough review of ratemaking systems for automobile insurance using cross-section data can be found in Denuit *et al.* (2007).

Nevertheless, in recent years, insurers have accumulated longitudinal information on their policyholders. To use this available panel or longitudinal data structure for ratemaking, the literature for panel count data models in the context of insurance has grown considerably in the past years. By using these models, actuaries may use repeated observations of each policyholder that are observed over time and, hence, they may avoid the classical two-step approach to ratemaking as Boucher and Inoussa (2014) stated. They concluded that the classical two-step approach are no longer coherent in a panel data setting and that the advantage of using the information oneach policyholder over time for modeling the number of claims is that we may estimate premiums that depend on risk characteristics and on claim experience simultaneously.

Following Molenberghs and Verbeke (2005), models for discrete panel data may be classified into three categories: conditional models (e.g. autoregressive and integer-valued autoregressive models), marginal models (e.g. multivariate models with serial correlation) and subject-specific models (e.g. random effects models). In the actuarial context, an exhaustive overview of such a models are presented in Boucher *et al.* (2008). They introduced some of these models to actuarial science and compared the fitting to the data. They concluded that the random effects models have a better fit than the other models. Other actuarial references that deserve to be highlighted are mentioned below. Gourieroux and Jasiak (2004) discussed the application of integer valued autoregressive models (INAR) for automobile insurance ratemaking. A comparative survey on random effects models for panel count data was conducted by Boucher and Guillén (2009).

In addition, when actuaries face the problem of pricing an insurance contract that contains different types of coverage, such as a motor insurance or homeowner's insurance policy, they usually assume that types of claim are independent. However, this assumption may not be realistic. Bermúdez (2009) and Bermúdez and Karlis

(2011, 2012) have shown that there is a positive correlation between types of claim and they introduced different bivariate (or multivariate) regression models in order to relax the independence assumption between claims counts arising from the same policy in an a priori ratemaking and a cross-section data setting. They concluded that using a bivariate (or multivariate) regression model leads to an a priori ratemaking that presents larger variances and, hence, larger loadings than those obtained under the independence assumption.

To sum up, Gourieroux and Jasiak (2004) and Boucher *et al*. (2008) showed that integer valued autoregressive models (INAR) are an acceptable alternative to model univariate claim count data when a panel data structure is available for ratemaking. Bermúdez (2009) and Bermúdez and Karlis (2011, 2012) showed that the use of bivariate (or multivariate) regression models for cross-section data when the ratemaking consists on pricing different types of coverage provides a better fit than the univariate regression models assuming independence. The present paper combines these two approaches and extends the integer valued autoregressive models (INAR) for panel claim count data to the bivariate case.

In this paper, a bivariate INAR(1) model introduced by Pedeli and Karlis (2011) is adapted to the ratemaking problem of pricing an automobile insurance contract with two types of coverages (third-party liability guarantee and other guarantees) taking into account both the correlation between claims from different type of coverage and the serial correlation between the observations of a same policyholder that are observed over time. So far little attention has been given to multivariate longitudinal data analysis for actuarial applications (Shi, 2012).

In the next section, the BINAR(1) model used here is defined. In Section 3, a numerical application using an automobile insurance claims database is presented. Finally, some concluding remarks are given in Section 4.

2. BINAR(1) MODEL

The bivariate INAR model is a generalization of the simple INAR model introduced by Al-Osh and Alzaid (1987) based on thinning.

Definition A sequence of random variables $\{X_t : t = 0, 1, \ldots\}$ is an INAR(1) process if it satisfies a difference equation of the form

$$X_t = \alpha \circ X_{t-1} + R_t; \quad t = 1, 2, \ldots$$

where $\alpha \in [0, 1]$, R_t is a sequence of uncorrelated non–negative integer– valued random variables with mean λ and finite variance σ^2 (called hereafter as the innovations) and X_0 represents an initial value of the process. The operator '$\circ$' is defined as

$$\alpha \circ X = \sum_{i=1}^{X} Z_i = Z$$

where Z_i are independently and identically distributed Bernoulli random variables with $P(Z_i = 1) = 1 - P(Z_i = 0) = \alpha$. If $X = 0$ then $Z = 0$ as well. This operator, known as the binomial thinning operator, is due to Steutel and van Harn (1979) and mimics the scalar multiplication used for normal time series models so as to ensure that only integer values will occur.

The model has been generalized to more dimensions, i.e. considering more than one time series together. Let $\mathbf{X}$ and $\mathbf{R}$ be non-negative integer-valued random 2-vectors. Let A be a 2×2 with independent elements $\{\alpha_{ij}\}_{i,j=1,2}$. The bivariate integer-valued autoregressive process of order 1 can be defined as

$$\mathbf{X}_t = \mathbf{A} \circ \mathbf{X}_{t-1} + \mathbf{R}_t$$

The model was described in detail in Pedeli and Karlis (2011) where the case with a diagonal matrix A and specific distributions for the vector R was examined. Application of the thinning operation in matrices implies element wise application of them in a similar fashion as matrix multiplication. In this note we focus on the case of diagonal A, since this has simpler structure. In this case each series is represented as $X_{jt} = \alpha_{jj} \circ X_{j,t-1} + R_{jt}$. The correlation comes from the joint distribution assumed for the R_{jt}.

Assuming independence between and within the thinning operations and $\{R_{jt}\}$ an *iid* sequence with mean λ_j and variance $\sigma_j^2 = \upsilon_j \lambda_j$, $\upsilon_j > 0, j = 1, 2$, the unconditional first and second order moments can be defined as

$$E(X_{jt}) = \mu_{X_j} = \frac{\lambda_j}{1-\alpha_{jj}}$$

$$Var(X_{jt}) = \sigma^2_{X_j} = \frac{(\alpha_{jj}+v_j)\lambda_j}{1-\alpha^2_{jj}}$$

It also holds that

$$corr\,(X_{jt}, X_{jt+h}) = \rho_{X_j}(h) = \alpha^h_{jj} \quad ; h \geq 1$$

i.e. the correlation decays exponentially as in an AR(1) process. More details about the model can be found in Pedeli and Karlis (2011) as well as extensions. The non-diagonal model has been considered in Boudreault and Charpentier (2011) and Pedeli and Karlis (2013).

3. ACTUARIAL APPLICATION

Suppose that for each individual we have two distinct type of claims and data for different time points. We denote as N_{1it} and N_{2it} the number of claims for the i-th individual at time point t, where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$ (i.e. we may have different number of observations for each client). In order to model such data we assume that for each individual we have observed a bivariate time series of BINAR(1) type. Also covariates can be available (however not treated in this paper). We aim to capture both the time series and the cross correlation aspects of the data. Hence we assume that

$$\begin{aligned} N_{1it} &= \alpha_{1i} \circ N_{1i,t-1} + R_{1it} \\ N_{2it} &= \alpha_{2i} \circ N_{2i,t-1} + R_{2it} \end{aligned}$$

were (R_{1it}, R_{2it}) are the innovation terms that follow jointly a bivariate Poisson distribution with parameters λ_{1it}, λ_{2it} and λ_{13t}. Finally note that parameters α and λ's are defined as client dependent and hence they can be related to some covariate information as well.

As in Pedeli and Karlis (2011) the conditional likelihood for a BINAR(1) model is the convolution of two binomials and a bivariate Poisson and hence the contribution of the i-th policyholder is

$$\begin{aligned} L_i &= \prod_{t=1}^{T_i} f(n_{1it}, n_{2it} | n_{1i,t-1}, n_{2i,t-1}, \alpha_{1i}, \alpha_{2i}, \lambda_{1it}, \lambda_{2it}, \lambda_{3it}) \\ &= \sum_{k=0}^{n_{1it}} \sum_{s=0}^{n_{2it}} f_1(n_{1it} - k) f_2(n_{2it} - s) f_3(k, s) \end{aligned}$$

with f_1 be the probability function of a binomial distribution with parameters $n_{1i,t-1}$ and α_{1i}, f_2 be the probability function of a binomial distribution with parameters $n_{2i,t-1}$ and α_{2i}, while $f_3(\cdot, \cdot)$ is the probability function of a bivariate Poisson distribution with parameters λ_{1it}, λ_{2it}, λ_{13t}.

Based on all these the full likelihood is just the product of the likelihood for each client, i.e.

$$\ell(\boldsymbol{\Theta}) = \prod_{i=1}^{n} L_i$$

Maximization is possible via standard numerical optimization. We have used R code to fit the model to the data.

Model	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	Log-lik
No cross correlation	0.042	0.168	0.0667	0.0685		-66823.77
No time series correlation			0.0566	0.0680	0.01393	-66789.14
BINAR model	0.040	0.162	0.0535	0.0557	0.0133	-65057.85

Table 1. Fitting different models to the data

The data used in this note are part of an automobile portfolio of a major insurance company operating in Spain in 1995. The data have been used also in Bermudez (2009). Only cars categorized as being for private use were considered. The data contains information from 20,000 randomly selected policy holders. We make use of 7 years of data for each policy holder. Hence for each individual we have 7 observations in

successive time points. Covariate information is also available but not treated in this note. Two type of claims are considered: N_1 is the number of claims for third-party liability and N_2 is number of claims for the rest of guarantees.

We have fitted the BINAR(1) model considered above together with some other models, namely a model that assumes no time series correlation (and hence considers the data as bivariate Poisson observations) and a model where there is not correlation between the two type of claims and hence fits two independent INAR(1) models (as the models in Boucher and Guillen, 2009). Results can be seen in Table 1. The estimated parameters and the log-likelihood of the fitted models are presented.

One can see that the improvement of the BINAR model over the simplest models is very large. This implies that we need to consider both time and cross correlation to fit the data at hand.

4. CONCLUSION

In this paper we fitted a longitudinal BINAR(1) model to data referring to the history of policyholders from an insurance company with two types of coverage(third-party liability guarantee and other guarantees). Such a model takes into account both the correlation between claims from different type of coverage and the serial correlation between the observations of a same policyholder. The implications of such a model, including covariates, to the ratemaking problem of pricing an automobile insurance contract is under progress.

REFERENCES

Al-Osh, M., and Alzaid, A. (1987). "First-Order Integer-Valued Autoregressive Process". *Journal of Time Series Analysis*, 8(3), 261-275.

Bermúdez, L. (2009). "A priori Ratemaking Using Bivariate Poisson Regression Models". *Insurance: Mathematics and Economics*, 44(1), 135-141.

Bermúdez, L., and Karlis, D. (2011). "Bayesian Multivariate Poisson Models for Insurance Ratemaking". *Insurance: Mathematics and Economics,* 48(2), 226-236.

Bermúdez, L., and Karlis, D. (2012). "A Finite Mixture of Bivariate Poisson Regression Models with an Application to Insurance Ratemaking". *Computational Statistics & Data Analysis,* 56, 3988-3999.

Boucher, J.-P., Denuit, M., and Guillén, M. (2008). "Models of Insurance Claim Counts with Time Dependence Based on Generalisation of Poisson and Negative Binomial Distributions", *Variance,* 2(1), 135-162.

Boucher, J.-P., and Guillén, M. (2009). "A Survey on Models for Panel Count Data with Applications to Insurance". *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, 103(2), 277-295.

Boucher, J.-P., and Inoussa, R. (2014). "A Posteriori Ratemaking with Panel Data". *ASTIN Bulletin,* 44, 587-612.

Boudreault, M., and Charpentier, A. (2011). *Multivariate Integer-Valued Autoregressive Models Applied to Earthquake Counts*. http://arxiv.org/abs/1112.0929.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.F. (2007). *Actuarial modelling of claim counts.* London: John Wiley & Sons.

Gourieroux, C., and Jasiak J. (2004). "Heterogeneous INAR(1) Model with Application to Car Insurance". *Insurance: Mathematics and Economics*, 34, 177-192.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Pedeli, X., and Karlis, D. (2011). "A Bivariate INAR(1) Process with Application". *Statistical Modelling,* 11(4), 327-351.

Pedeli, X., and Karlis, D. (2013). "Some Properties of Multivariate INAR(1) processes". *Computational Statistics and Data Analysis,* 67, 213-225.

Shi, P. (2012). "Multivariate Longitudinal Modeling of Insurance Company Expenses". *Insurance: Mathematics and Economics*, 51(1), 204-215.

Steutel, F., and Van Harn, K. (1979). "Discrete Analogues of Self-Decomposability and Stability". *The Annals of Probability,* 7, 893-899.

17. A FINITE MIXTURE OF MULTIPLE POISSON DISTRIBUTIONS FOR MODELLING DAYS OF ABSENCE FROM WORK DUE TO MOTOR ACCIDENT

L. Bermúdez
Universitat de Barcelona, Riskcenter-IREA, Departament de Matemàtica Econòmica, Financera i Actuarial, Spain, lbermudez@ub.edu

D. Karlis
Athens University of Economics and Business, Department of Statistics, Greece, karlis@aueb.gr

M. Santolino
Universitat de Barcelona, Riskcenter-IREA, Departament d'Estadística, Econometria i Economia Espanyola, Spain, msantolino@ub.edu

ABSTRACT

In this paper we present a model to fit appropriately data with a lot of spikes in certain values. A new model, based on finite mixtures of multiple Poisson distributions of different multiplicities, is proposed to fit this kind of data. A numerical application with a real dataset concerning the length, measured in days, of inability to work after an accident occurs is treated. The main finding is that the model provides a very good fit when working week, calendar week and month multiplicities are taken into account.

1. INTRODUCTION

This paper aims at presenting a model suitable to fit data with a lot of spikes in certain values. In particular, we work with data that refer to the length, measured in

days, of inability to work after an accident occurs. In this case, the frequency distribution exhibits spikes at multiples of 5, 7 and 30 days, implying perhaps the different time scales used by doctors when deciding on the number of days for leave to the workers.

We emphasize that such spikes can occur in certain other applications, like digit preference, declared age in demography (also known as heaping), etc. For example, Camarda *et al.* (2008) used a model where probabilities are moved to the spike points by the adjacent ones. However, our approach is very different.

A new model based on finite mixtures of multiple Poisson distributions of different multiplicities is proposed to fit this kind of data. This model is described in Section 2. In Section 3 an EM algorithm is proposed for maximum likelihood estimation. A numerical application with a real dataset of days of absence is presented in Section 4. Finally, some concluding remarks are given in Section 5.

2. THE MODEL

First of all, consider a simple Poisson distribution with parameter λ denoted as $Po(\lambda)$ with probability function (pf) given by

$$P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \lambda \geq 0, x = 0, 1, 2, \ldots$$

We extend by defining the multiple Poisson, with multiplicity m and parameter λ, denoted as $Pm\,(\lambda)$ having pf

$$P_m(X = y) = \begin{cases} \frac{\exp(-\lambda)\lambda^x}{x!} & y = mx, \ x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Hence this distribution gives positive probability to points $0, m, 2m, \ldots$ and 0 elsewhere. The idea for this representation is that we can have data measured in different scales, as for example observations measured in days, weeks and months. In this

case, $P_1(\lambda_1)$, $P_7(\lambda_7)$ and $P_{30}(\lambda_{30})$ respectively. The derivation of the above model is not new in the sense that it can be seen as an special case of what is known as *stuttering Poisson* model (see, Patel, 1976 for example). However, our definitive modelling approach, based on the above model and explained bellow, is new up to our knowledge. Following the above multiple Poisson distribution with parameters m and λ, we got that

$$E(X) = m\lambda \text{ and } Var(X) = m^2\lambda.$$

Also, the probability generating function takes the form

$$G(t) = \exp\left(\lambda(t^m - 1)\right).$$

The idea is to construct a new model based on finite mixtures of multiple Poisson distributions of different multiplicities. This, for example, can be attributed to the fact that different scales are used implicitly. For example, a doctor based on the severity of some injury may think on a daily or weekly or monthly scale for the number of days of absence for a patient. It is rather unprobable to say 13 days of absense but more probable to say 14 days (2 weeks). Hence the observed count can have spikes due to different scales. Unfortunately the original scale is not kept and hence we have only the observations at hand without knowledge on the scale. This leads naturally to a mixture and typical such data can have a lot of spikes.

Denote by Y_i the count of some event for $i = 1, \ldots, n$ observations. Conditional on the fact that the i-th observation belongs to the j-th group, the observed counts come from a Poisson distribution with mean λ_j and multiplicity m_j. The probability that a randomly selected individual belongs to group j is $0 < \pi_j < 1$ for $j = 1, \ldots, k$ with $\Sigma\, \pi_j = 1$. Then the unconditional pdf is

$$P(Y_i = y) = P(y) = \sum_{i=1}^{K} \pi_j P_{m_j}(y|\lambda_j), \quad y = 0, 1, \ldots$$

where K is the number of groups, λ_j are the means for each group, m_j are the multiplicities. This is a typical finite mixture model. This implies that many of the ideas developed so far for finite mixtures are applicable here, like derivation of an EM type algorithm, selection of the number of components, etc.

This simple model can be extended if we allow the mean of each component to depend on some covariates related to the i-th individual but this is not pursued in the present paper. Also in the present paper, we will work with multiple Poisson models but clearly extension to other models, i.e. negative binomial, is possible.

3. AN EM ALGORITHM

In this section we describe briefly an EM algorithm for ML estimation. Before doing so consider the simple case when a random sample of size n from a P_m distribution is available. Then the ML estimate of λ is simply

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{mn}$$

Note that if the data come from the Pm then the integer division will be valid in all cases.

Consider the case of K components, with multiplicities mj and parameters λj. Note that m_j are known and thus we do not need to estimate them. Using the standard approach for finite mixtures we assume the unobserved latent variables Z_{ij} which are the component indicator i.e. $Z_{ij} = 1$ if the i-th observation belongs to the j-th group and 0 otherwise. Hence $Z_i = (Z_{il}, \ldots, Z_{ik})$. Note that for our model and a specific multiplicity some observations clearly have $Z_{ij} = 0$ by definition. For example for multiplicity equal to 2 every odd observations cannot be generated from this component. Then the steps of the algorithm are the following:

At the E-step using the current estimates calculate

$$w_{ij} = \frac{\pi_j P_{m_j}(y_i|\lambda_j)}{P(y_i)}$$

and then update the parameters at the M-step as

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n} \quad \text{and} \quad \hat{\lambda}_j = \frac{\sum_{i=1}^{n} w_{ij} x_i}{m_j \sum_{i=1}^{n} w_{ij}}$$

All the advantage and disadvantages of the EM for finite mixtures apply here too. Of course typical cautions about EM apply. We need good initial values in order to avoid problems of local maxima. Also, convergence can be slow.

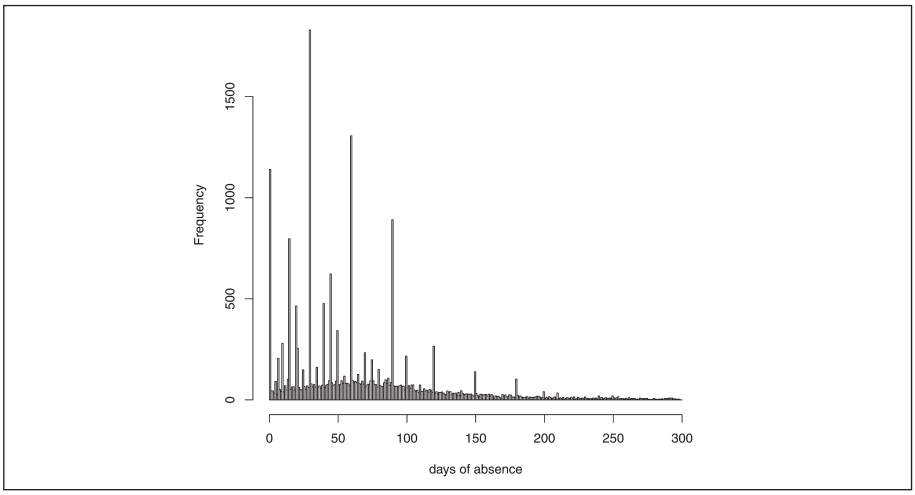


Figure 1. The observed frequencies for the data.

4. APPLICATION: DAYS OF ABSENCE

The data refer to the length of inability to work. We make use of 19960 observations. The most interesting feature of the data is that there are a lot of spikes in certain values. The data are depicted in Fig. 1. A close look on these values shows, for example, spikes in multiples of 5, 7 and 30, implying perhaps the different time scale used by doctors when deciding on the number of days for leave to the workers. For example, it is reasonable that a doctor advises re-examination of a patient after 30 days (one month) rather than 29 or 31 days. This creates the spikes in the frequency distribution which we aim at modelling with the model described in section 2.

From the plot one can see spikes at certain values. Obviously there are significant spikes on 5 days (a working week), 7 days (calendar week), 10 days, 15 days, 20 days, 30 days (month), 60 days and 90 days. Also the variance of the observed values is very large. The variance equals 3160 and the mean equals 68.22. This implies that even for the simple Poisson case (i.e. if $m = 1$) we need more components to fit this

part. The same is perhaps valid for other values of m, where more than one component would be needed. Also note that all components contribute to the zero cell which is very large as one can see in Fig. 1.

In order to model this dataset we assumed a mixture of multiple Poisson distributions as described in section 2. Table 1 reports the estimated components. We used 19 components for this data set. The first 11 had $m = 1$ and well separated components that try to fit the general picture. This perhaps implies that this component can be replaced from an overdispersed distribution like the negative binomial rather than many simple Poisson ones. The rest components fit the spikes present. We also report at the last column of the Table the log-likelihood when the specific components was removed as an indication how important is this component. The log-likelihood of the full model is -91265.74. One can see some interesting features. For all values of m if we remove this component the log-likelihood decreases a lot, implying that the existence of this component is very important for the model. Only for $m = 5$ we need two components to fit this part. While we have tried with the other multiplicities the results coincide with just using one component for this multiplicity.

To have an idea on the goodness of fit we have calculated a chi-square goodness of fit statistic. Cells with expected frequency lower than 5 were removed from the calculations. In total 275 out of 300 cells were used, and the sum of the expected frequencies not used was only 91 (out of 19960). The chi-square test statistic was 27.63 which implies a very good fit for 234 degrees of freedom. Finally, Fig. 2 shows the difference between observed and expected frequencies. One can see as checked with 2 that the fit is very good in a great extend.

m	λ	π	Log-lik if removed
1	2.8476	0.0090	
1	9.9614	0.0203	
1	20.8443	0.0394	
1	35.8674	0.0540	
1	52.7750	0.0767	
1	72.0698	0.0887	
1	96.3278	0.0947	
1	126.8449	0.0556	
1	161.7984	0.0394	
1	204.1400	0.0268	
1	260.6124	0.0212	
5	4.0235	0.0352	-91372.29
5	16.3175	0.0177	
7	2.1336	0.0262	-91460.22
10	4.0051	0.0888	-91398.46
15	2.1845	0.1302	-91781.10
20	5.6823	0.0119	-91292.23
30	2.0783	0.1318	-91674.76
90	0.4533	0.0324	-91325.08

Table 1. The fitted components with their estimated parameters.

For all estimations we used the EM algorithm. The stopping criterion was to stop when the relative change in the log-likelihood between two successive iterations was smaller than 10–8. As initial values we used an incremental approach, where each new component was given probability 0.05 (decreasing the probability of the existing components), with mean at the center (i.e. that we would obtain if only this components was used). Decisions on whether to add a new multiplicity was decided by inspection of the residuals, i.e. the difference between the observed and the fitted frequencies of the current model.

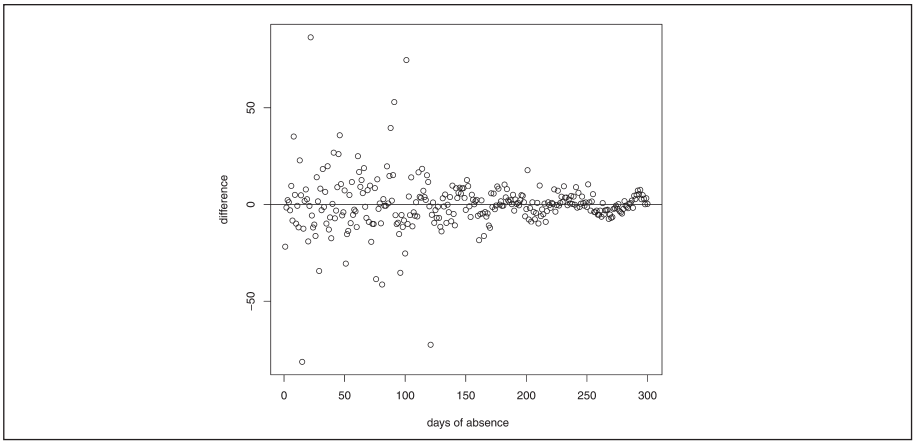


Figure 2. Differences between observed and expected frequencies from the model.

5. CONCLUSION

In this paper we developed a new model for modelling the days of absence. A specific feature of such data is the presence of spikes in certain number of days perhaps due to different calendar scales that are used by medical doctors depending on the severity of the injury. An EM algorithm was used to fit the model to a real dataset. Extension to the case when covariates are present is straightforward but we do not report it in this paper. Finally, while in this paper we worked only the case of multiple Poisson model extension to other distributions is simple.

REFERENCES

Camarda, C.G., Eilers, P.H.C., and Gampe, J. (2008). "Modelling General Patterns of Digit Preference", *Statistical Modelling,* 8(4), 385-401.

Patel, Y.C. (1976). "Estimation of the Parameters of the Triple and Quadruple Stuttering-Poisson Distributions". *Technometrics*, 18, 67-73.

18. INFLUENCIA DEL REASEGURO Y DE LA ESTRUCTURA DEL COLECTIVO EN LA SOLVENCIA DE UNA CARTERA DE VIDA

E. Boira
Vida Caixa, Área de Desarrollo de Negocio, España,
eboira@vidacaixa.es

M.A. Pons
Universidad de Barcelona, Departamento de Matemática Económica, Financiera y Actuarial, España,
mapons@ub.edu

F.J. Sarrasí
Universidad de Barcelona, Departamento de Matemática Económica, Financiera y Actuarial, España,
sarrasi@ub.edu

RESUMEN

El objetivo de este trabajo es obtener, desde el punto de vista de la compañía de seguros, y fijada su política de reaseguro, el capital de solvencia, que garantiza una probabilidad de solvencia, asociado a una cartera del ramo de vida constituida por seguros de fallecimiento anuales. Para su cálculo es necesario determinar la distribución de probabilidad de la variable aleatoria coste total de la cartera retenida por la compañía de seguros, la cual se obtiene estimando la evolución de la siniestralidad de dicha cartera a través del método de simulación de Monte Carlo. Por último, se realiza un análisis de cómo afecta al ratio de solvencia de la compañía su política de reaseguro.

1. INTRODUCCIÓN

El trabajo se centra en el estudio y análisis del capital de solvencia que debe tener dotado una compañía de seguros para garantizar una probabilidad de solvencia ε, asociado a una cartera de seguros de fallecimiento anuales, teniendo en cuenta la política de reaseguro de la compañía. En particular, las modalidades de reaseguro consideradas son el reaseguro cuota parte y el reaseguro de excedentes. Dada la cartera de la cedente, se simula la evolución anual de la siniestralidad de la misma mediante el método de Monte Carlo. Una vez obtenida la función de distribución de la variable aleatoria simulada, evolución de la cartera en el transcurso de un año, se obtiene la distribución de probabilidad de la variable aleatoria coste total de la cartera de la cedente y, a partir de ella, su esperanza, su desviación típica y su percentil o Var al 99,5% teniendo en cuenta las dos modalidades de reaseguro consideradas. El capital de solvencia, que garantiza una solvencia del 99,5% en el horizonte temporal de un año, se obtiene mediante el Var al 99,5% de la variable aleatoria coste total.

2. METODOLOGÍA

Dada una cartera N formada por un colectivo de n asegurados, beneficiarios de un seguro de fallecimiento anual, se define T_N como la variable aleatoria evolución de la cartera N en el transcurso de un año, siendo:

$$T_N = (T_{x_1}, T_{x_2}, \ldots, T_{x_r}, \ldots, T_{x_n}),$$

donde T_{xr}, con $r = 1, \ldots, n$, es la variable aleatoria dicotómica fallecimiento del asegurado de r edad x_r en el transcurso de un año, cuya distribución de probabilidad se muestra en la Tabla 1, donde T_{xr} es la realización de la variable aleatoria T_{xr}, con $r = 1, \ldots, n$.

t_{x_r}	$P[T_{x_r} = t_{x_r}]$
1	q_{x_r}
0	$1 - q_{x_r}$

Tabla 1. Distribución de probabilidad de T_{xr}.

En la Tabla 1, el cero indica la supervivencia del asegurado y el uno, el fallecimiento del mismo en el transcurso de un año. Por tanto, las realizaciones de la variable T_N son vectores, $\overrightarrow{t_N}$ que tienen tantas componentes como asegurados tiene la cartera N:

$$\overrightarrow{t_N} = (t_{x_1}, t_{x_2}, \dots, t_{x_r}, \dots, t_{x_n}).$$

En la Tabla 2 se muestra la función de distribución de la variable aleatoria T_N.

$\overrightarrow{t_N}$	$P[T_N = \overrightarrow{t_N}]$
(0,0,0,...,0)	$(1-q_{x_1})\cdot(1-q_{x_2})\cdot(1-q_{x_3})\cdot\dots\cdot(1-q_{x_n})$
(1,0,0,...,0)	$q_{x_1}\cdot(1-q_{x_2})\cdot(1-q_{x_3})\cdot\dots\cdot(1-q_{x_n})$
...	...
(1,1,1,...,1)	$q_{x_1}\cdot q_{x_2}\cdot q_{x_3}\cdot\dots\cdot q_{x_n}$

Tabla 2. Distribución de probabilidad de T_N.

Al ser 2^n el número de realizaciones de la variable T_N, prácticamente se hace imposible trabajar directamente con su función de distribución, a no ser que se trate de un colectivo muy pequeño. Para solucionar este problema, se simularán, por el método de Monte Carlo, las trayectorias de siniestralidad de la cartera. El objetivo de la simulación es obtener la función de distribución de la variable aleatoria T_N a partir de la función de distribución de su variable aleatoria simulada, $\widetilde{T_N}$.

El número de realizaciones de $\widetilde{T_N}$ ya no será 2^n , sino el número de simulaciones que se realicen. Si α es el número total de simulaciones efectuadas en la cartera, las realizaciones, $\widetilde{T_N}$, de la variable aleatoria simulada (vienen dadas por el siguiente conjunto discreto:

$$\widetilde{t_N} = \{\overrightarrow{t^1}, \overrightarrow{t^2}, \dots, \overrightarrow{t^z}, \dots, \overrightarrow{t^\alpha}\},$$

siendo $\overrightarrow{t^z} = (\widetilde{t^z_{x_1}}, \widetilde{t^z_{x_2}}, \dots, \widetilde{t^z_{x_r}}, \dots, \widetilde{t^z_{x_n}})$, con $z = 1,2, \dots, \alpha)$, y donde , con $r = 1,2, \dots, n$, es la variable aleatoria dicotómica fallecimiento del asegurado r de edad x_r en el transcurso de un año, asociada a la simulación z-ésima.

Simular $\widetilde{t_{x_r}^z}$ por Monte Carlo supone generar un número aleatorio $U \sim U(0,1)$, de tal forma que:

$$\widetilde{t_{x_r}^z} = \begin{cases} 1 & si \quad U \leq q_{x_r}. \\ 0 & si \quad U > q_{x_r}. \end{cases}$$

Si se repite este proceso de simulación α veces, se obtiene la función de distribución de la variable aleatoria $\widetilde{T_N}$, que se muestra en la Tabla 3.

$\widetilde{t_N}$	$P[\widetilde{T_N} = \widetilde{t_N}]$	Probabilidad acumulada
$\overrightarrow{t^1}$	$1/\alpha$	$1/\alpha$
$\overrightarrow{t^2}$	$1/\alpha$	$2/\alpha$
...	...	...
$\overrightarrow{t^\alpha}$	$1/\alpha$	1

Tabla 3. Distribución de probabilidad de $\widetilde{T_N}$.

Cada realización de la variable aleatoria $\widetilde{T_N}$ supone una determinada trayectoria de evolución de la siniestralidad de la cartera N, y hay tantas trayectorias como simulaciones se hagan de la misma. La probabilidad asociada a cada realización de $\widetilde{T_N}$ es siempre la misma, ya que éstas se obtienen a partir de una distribución uniforme y, por tanto, su valor dependerá del número de simulaciones realizadas.

Conocida la función de distribución de $\widetilde{T_N}$, se puede obtener la función de distribución de la variable aleatoria coste total de la cartera de la cedente, $\widetilde{V_N}$. Su aleatoriedad viene inducida por la variable aleatoria simulada $\widetilde{T_N}$, ya que el coste total de la cartera se obtiene como el producto:

$$\widetilde{V_N} = \overrightarrow{S_N^c} \times \widetilde{T_N},$$

donde $\overrightarrow{S_N^c}$ es el vector de sumas aseguradas a cargo de la cedente asociada a la cartera N:

$$\overrightarrow{S_N^c} = (S_1^c, S_2^c, \dots, S_r^c, \dots, S_n^c),$$

siendo S_r^c, con $r = 1, \dots, n$, la suma asegurada a cargo de la cedente asociada al asegurado r-ésimo de edad actuarial x_r.

El valor de S_r^c, con $r = 1, \dots, n$, dependerá de la modalidad de reaseguro que tenga contratada la cedente:

- Si el reaseguro es cuota parte:

$$S_r^c = S_r \cdot k \quad \text{con} \quad r = 1,2, \dots, n,$$

siendo S_r la suma asegurada asociada al asegurado r-ésimo y k, con $0 < k \leq 1$, la cuota de retención de la cedente en tanto por uno. En el caso particular de que $k = 1$, la cedente estaría reteniendo la totalidad de la cartera.

- Si el reaseguro es de excedentes:

$$S_r^c = \begin{cases} S_r & si \quad S_r \leq M \\ M & si \quad S_r > M \end{cases} \quad \text{con} \quad r = 1,2, \dots, n,$$

siendo M el pleno de retención de la cedente.

La variable aleatoria $\widetilde{V_N}$ tendrá tantas realizaciones como simulaciones se realicen de las trayectorias de siniestralidad de la cartera N. Las realizaciones, $\widetilde{v_N}$, de la variable aleatoria $\widetilde{V_N}$ siguiente conjunto discreto:

$$\widetilde{v_N} = \{\widetilde{v_1}, \widetilde{v_2}, \dots \widetilde{v_z} \dots \widetilde{v_\alpha}\},$$

siendo $\widetilde{v_z}$, con $z = 1, \dots, \alpha$), el valor obtenido como el producto escalar entre el vector $\overrightarrow{S_N^c}$ y el vector $\overrightarrow{t^z}$:

$$\widetilde{v_z} = \overrightarrow{S_N^c} \times \overrightarrow{t^z} = \sum_{r=1}^{n} S_r^c \cdot \widetilde{t_{x_r}^z}.$$

La función de distribución de la variable $\widetilde{V_N}$ se muestra en la Tabla 4.

$\widetilde{v_N}$	$P[\widetilde{V_N} = \widetilde{v_N}]$	Probabilidad acumulada
$\widetilde{v_1}$	$1/\alpha$	$1/\alpha$
$\widetilde{v_2}$	$1/\alpha$	$2/\alpha$
…	…	…
$\widetilde{v_z}$	$1/\alpha$	z/α
…	…	…
$\widetilde{v_\alpha}$	$1/\alpha$	1

Tabla 4. Distribución de probabilidad de $\widetilde{V_N}$.

Conocida la función de distribución de la variable $\widetilde{V_N}$, se pueden calcularlos estadísticos más representativos, como su esperanza, $E[\widetilde{V_N}]$, su desviación típica, $VaR_\varepsilon(\widetilde{V_N})$, y su percentil o $VaR_\varepsilon(\widetilde{V_N})$ asociado a un nivel de solvencia Ɛ, siendo respectivamente:

$$E\left[\widetilde{V_N}\right] = \frac{\sum_{z=1}^{\alpha} \widetilde{v_z}}{\alpha}, \quad D\left[\widetilde{V_N}\right] = \sqrt{E\left[\widetilde{V_N}^2\right] - E\left[\widetilde{V_N}\right]^2} \text{ y } P\left[\widetilde{V_N} \geq VaR_\varepsilon\left(\widetilde{V_N}\right)\right] = 1 - \varepsilon.$$

El capital de solvencia que garantiza una probabilidad de solvencia ε viene dado por el $VaR_\varepsilon(\widetilde{V_N})$. A partir del capital de solvencia se define el ratio de solvencia de la cedente asociado a una probabilidad de solvencia ε, τ_ε, como el cociente entre sus recursos propios y el $VaR_\varepsilon(\widetilde{V_N})$. Los recursos propios vienen dados por las reservas de solvencia, R, que la cedente tiene disponibles en el plazo considerado, y por las primas asociadas a la cartera N, cuyo importe viene dado, si se asume el criterio de la esperanza matemática para su cálculo, por $E\left[\widetilde{V_N}\right] \cdot (1 + \delta)$, siendo δ el recargo de seguridad:

$$\tau_\varepsilon = \frac{E[\widetilde{V_N}] \cdot (1+\delta) + R}{VaR_\varepsilon(\widetilde{V_N})}.$$

3. APLICACIÓN NUMÉRICA

La cartera N objeto de estudio está constituida por un colectivo formado por n = 276 asegurados, beneficiarios de un seguro de fallecimiento anual. La estructura de edades y de prestaciones vienen reflejadas en la Tabla 5.

Número de asegurados	Edad	Suma asegurada
51	21	3.000
39	34	3.500
42	40	3.700
42	47	4.000
30	50	4.000
36	65	4.300
36	70	4.500

Tabla 5. Estructura de la cartera N.

Los valores que se muestran a continuación han sido obtenidos en FORTRAN tras realizar 1.000.000 de simulaciones de la cartera N y utilizando como bases técnicas las tablas de mortalidad PASEM 2010.

En la Tabla 6 se muestra, para un reaseguro cuota parte y excedentes, los valores de $E[\widetilde{V_N}]$, $D[\widetilde{V_N}]$, $VaR_{99,5\%}[\widetilde{V_N}]$ y $\tau_{99,5\%}$ para un recargo de seguridad $\delta = 0{,}05$ y unas reservas de solvencia $R = 10.000$.

	k = 0,5	k = 0,75	k = 1	M = = 3.000	M = = 3.500	M = = 4.000
$E[\widetilde{V_N}]$	3.493,09	5.239,63	6.986,18	4.871,30	5.675,03	6.445,43
$D[\widetilde{V_N}]$	2.723,02	4.084,53	5.446,04	3.791,62	4.417,50	5.018,80
$VaR_{\varepsilon}(\widetilde{V_N})_{99,5\%}$	12.650	19.975	25.300	18.000	21.000	23.700
$\tau_{99,5\%}$	1,080	0,8169	0,6851	0,8397	0,7599	0,7074

Tabla 6. Resultados de reaseguro cuota parte y excedente.

Cuanto mayor es la cartera cedida al reasegurador por parte de la cedente, mayor es su ratio de solvencia.

En la Tabla 7 se muestra, para $\delta = 0{,}05$, el valor de las reservas de solvencia, R, que garantiza un ratio de solvencia $\tau_{99,5\%} = 1$. En este caso:

$$R = VaR_{\varepsilon}\left(\widetilde{V_N}\right) - E\left[\widetilde{V_N}\right] \cdot (1+\delta) \quad \text{ya que} \quad \frac{E[\widetilde{V_N}]\cdot(1+\delta)+R}{VaR_{\varepsilon}(\widetilde{V_N})} = 1.$$

Reaseguro cuota parte	$k = 0{,}5$	$k = 0{,}75$	$k = 1$
R	8.982	13.473	17.964
Reaseguro de excedentes	$M = 3.000$	$M = 3.500$	$M = 4.000$
R	12.885,13	15.041	16.932,29

Tabla 7. Valor de las reservas de solvencia, R.

El valor de las reservas de solvencia R obtenido en la tabla anterior es el que tiene que dotar la cedente para garantizar un capital de solvencia del 99,5%. Su importe disminuye cuanto mayor es la cartera cedida al reaseguro.

4. CONCLUSIÓN

El capital de solvencia de una compañía de seguros viene dado por el $VaR_\varepsilon(\widetilde{V_N})$ de la variable aleatoria coste total, y su valor depende del nivel de solvencia deseado, ε, de la política de reaseguro y de la estructura de edades y de prestaciones de la cartera N. El ratio de solvencia de la cedente, obtenido como cociente entre sus recursos propios y el $VaR_\varepsilon(\widetilde{V_N})$, aumenta cuanto mayor es la cartera cedida al reasegurador.

BIBLIOGRAFÍA

Benito Rivero, J. (2001). *El reaseguro*. Editorial MAPFRE, S.A. Madrid.

Korn, R., Korn, E., y G. Kroisandt (2010). *Monte Carlo Methods and Models in Finance and Insurance*. Taylor and Francis Group. Londres.

Rodríguez Mondragón, J. (2009). *Optimización económica del reaseguro cedido: Modelos de decisión*. Fundación MAPFRE. Madrid.

Rusell M., y Collins, J. (1962). "Actuarial Aplication of the Monte Carlo Tecnique". *Transactions of the Society of Actuaries*. Vol. 14. No. 40.

19. CLAIM RESERVING: CALENDAR YEAR RESERVES FOR THE GLM

E. Boj
Universitat de Barcelona, Departament de Matemàtica Econòmica Financera i Actuarial, Facultat d'Economia i Empresa, Spain,
evaboj@ub.edu

T. Costa
Universitat de Barcelona, Departament de Matemàtica Econòmica Financera i Actuarial, Facultat d'Economia i Empresa, Spain,
tcosta@ub.edu

ABSTRACT

We obtain the formulas of the prediction error for the calendar year IBNR reserves when we assume a parametric family for the error distribution in the generalized linear model (GLM) which models the data in a run-off triangle. The parametric family has as particular cases the Poisson, the Gamma and the Inverse Gaussian distributions. We assume the logarithmic link function.

1. INTRODUCTION

We assume a GLM (see, e.g., Boj *et al*., 2004; Boj and Costa, 2014 and McCullagh and Nelder, 1989) to model a run-off triangle and the same nomenclature as in Boj *et al*. (2014). We assume the logarithmic link function $\log \mu_{ij} = \eta_{ij}$ and the parametric family of distributions:

$$\mu_{ij} = E\left[c_{ij}\right] \qquad Var\left[c_{ij}\right] = \phi V\left(\mu_{ij}\right) = \phi \mu_{ij}^{\theta}, \tag{1}$$

which depends on a parameter θ. When $\theta = 1$ we obtain the Poisson distribution; when $\theta = 2$ the Gamma distribution; and when $\theta = 3$ the Inverse Gaussian distribution. The prediction error of that model is studied for the origin year and total reserves in, e.g., England and Verrall (1999, 2002, 2006) and England (2002). The study is extended to the calendar year IBNR in Boj *et al.* (2014) and Espejo *et al.* (2014) but only for the particular case of the (overdispersed) Poisson distribution, which gives the same estimations of IBNR as the deterministic Chain-Ladder method. In the next section we propose the formulas for the calendar year reserves when we assume (1).

2. PREDICTION ERROR FOR CALENDAR YEAR RESEVES

First we show the formulas for the analytic deduction of the prediction error and second we show the formulas for the bootstrap estimation.

The prediction error in the analytical formula of the distribution can be calculated as the root of the mean squared error:

$$MSE\left(R_t\right) = E\left[\left(R_t - \hat{R}_t\right)^2\right] \approx \sum_{\substack{i,j=1,..,k \\ i+j=t}} \phi\hat{\mu}_{ij}^{\theta} + \mu_t^T Var\left[\eta_t\right]\mu_t = \tag{2}$$

$$= \sum_{\substack{i,j=1,..,k \\ i+j=t}} \phi\mu_{ij}^{\theta} + \sum_{\substack{i,j=1,..,k \\ i+j=t}} \mu_{ij}^2 Var\left[\eta_{ij}\right] + 2 \sum_{\substack{i_1,j_2,j_1,j_2=1,..,k \\ i_1,j_1 \neq i_2,j_2 \\ i_1+j_1=t, i_2+j_2=t}} \mu_{i_1,j_1}\mu_{i_2,j_2} Cov\left[\eta_{i_1,j_1}, \eta_{i_2,j_2}\right].$$

$$t = k+1, k+2, ..., 2k$$

The prediction error in the bootstrap estimation can be calculated as:

$$PE_{boot}(R_t) \cong \sqrt{\sum_{\substack{i,j=1,..,k \\ i+j=t}} \hat{\phi}^P \hat{\mu}_{ij}^{\theta} + SE_{boot}\left(R_t\right)^2}, \quad t = k+1, k+2, ..., 2k. \tag{3}$$

Note that in the bootstrapping residuals we use the Pearson residuals, $r_{ij}^P = \frac{c_{ij} - \hat{c}_{ij}}{\sqrt{\hat{c}_{ij}^{\theta}}}$, which we adjust for the degrees of freedom $r_{ij}^{P'} = \sqrt{\frac{n}{n-p}} r_{ij}^P$ in the same way as the scale parameter, $\hat{\phi}^P$, to be compatible in the bootstrap process. In the resamples we have *B* times $r_{ij}^{P'*}$, and with them we calculate $c_{ij}^* = r_{ij}^{P'*}\left(\hat{c}_{ij}^{\theta}\right)^{1/2} + \hat{c}_{ij}$.

3. APPLICATION

To illustrate computations we use the triangle of Taylor and Ashe (1983) with incremental losses. This dataset is used in many texts on IBNR problems as are England and Verrall (1999), England (2002) or Renshaw (1989, 1994). Specifically, in England and Verrall (1999) we can find the results for the origin year and total reserves for the overdispersed Poisson distribution using analytic formula and bootstrap methodology, and for the Gamma distribution only using analytic formula. In this section we complete the study for the calendar year reserves for the two distributions, the Poisson and the Gamma, using the two ways of calculating the prediction error, with analytic formula and with bootstrap using 1,000 resamples. And we compute the actual value assuming a 1.5% of annual interest rate in three cases: only to the IBNR values, to the IBNR values plus a percentage (25%) of the prediction error, and to the VaR at the confidence level of the 75%. We have done computations with the R software and we have fitted the models with the function *glm* of the *stats* package for R. In Tables 1 to 8 we show the results. We observe that, in general for this dataset, we obtain lower prediction errors for $\theta = 2$, the case of the Gamma distribution, if we compare the results with those of the overdispersed Poisson ($\theta = 1$).

Calendar year	Reserve	Prediction error	Coefficient of variation
10	5226535.8	416634.9	7.97 %
11	4179394.4	434848.9	10.40 %
12	3131667.5	462604.3	14.77 %
13	2127271.9	284150.8	13.36 %
14	1561878.9	217367.3	13.92 %
15	1177743.7	206110.4	17.50 %
16	744287.4	159987.2	21.50 %
17	445521.3	221521.6	49.72 %
18	86554.6	108268.8	125.09 %

Table 1. Calendar year reserves, prediction errors and coefficients of variation for the overdispersed Poisson distribution using analytic formula.

Calendar year	Reserve	Prediction error	Coefficient of variation
10	5096855.3	413065.7	8.10 %
11	4050001.5	424651.2	10.49 %
12	3064407.7	456009.8	14.88 %
13	2078010.5	256220.4	12.33 %
14	1510392.7	190750.1	12.63 %
15	1095402.7	163737.3	14.95 %
16	692118.4	112426.4	16.24 %
17	416539.9	172229.2	41.35 %
18	82075.9	47918.1	58.38 %

Table 2. Calendar year reserves, prediction errors and coefficients of variation for the Gamma distribution using analytic formula.

Calendar year	Mean reserve	Standard deviation	Prediction error	Coefficient of variation
10	5262187.6	748660.0	756563.2	14.48 %
11	4206004.9	718618.2	721067.3	17.25 %
12	3153556.8	653134.1	649753.2	20.75 %
13	2139244.8	504395.7	487995.9	22.94 %
14	1562523.4	408073.2	411005.7	26.31 %
15	1178586.1	364102.8	365547.7	31.04 %
16	771451.6	302919.9	292974.4	39.36 %
17	455633.0	250906.3	254458.2	57.11 %
18	91579.0	104329.2	107988.8	124.76 %

Table 3. Calendar year mean reserves, standard deviations, prediction errors and coefficients of variation for the overdispersed Poisson distribution using bootstrap methodology with 1,000 resamples.

Calendar year	Mean reserve	Standard deviation	Prediction error	Coefficient of variation
10	5096897.24	1017.5	652964.8	12.81 %
11	4050047.9	1014.2	545647.9	13.47 %
12	3064465.1	920.9	434825.7	14.19 %
13	2078021.4	694.1	297581.3	14.32 %
14	1510393.3	580.7	233914.7	15.49 %
15	1095419.1	493.1	194252.2	17.73 %
16	692130.4	395.9	142601.5	20.60 %
17	416548.9	342.7	109441.6	26.27 %
18	82080.9	152.7	26649.2	32.47 %

Table 4. Calendar year mean reserves, standard deviations, prediction errors and coefficients of variation for the Gamma distribution using bootstrap methodology with 1,000 resamples.

Calendar year	Deferral (in years)	Reserve	Reserve + 0.25 Prediction Error
10	1	5226535.8	5330694.6
11	2	4179394.4	4288106.7
12	3	3131667.5	3247318.6
13	4	2127271.9	2198309.6
14	5	1561878.9	1616220.7
15	6	1177743.7	1229271.3
16	7	744287.4	784284.2
17	8	445521.3	500901.7
18	9	86554.6	113621.8
Present Value		17873967	18466079

Table 5. Present values of the calendar year reserves and of the calendar year reserves plus a 25% of the prediction error for the overdispersed Poisson distribution using analytic formula.

Calendar year	Deferral (in years)	Reserve	Reserve + 0.25 Prediction Error
10	1	5096855.3	5200121.7
11	2	4050001.5	4156164.3
12	3	3064407.7	3178410.1
13	4	2078010.5	2142065.6
14	5	1510392.7	1558080.2
15	6	1095402.7	1136337.0
16	7	692118.4	720225.0
17	8	416539.9	459597.2
18	9	82075.9	94055.4
Present Value		17310125	17840014

Table 6. Present values of the calendar year reserves and of the calendar year reserves plus a 25% of the prediction error for the Gamma distribution using analytic formula.

Calendar year	Deferral (in years)	Reserve	Reserve + 0.25 Prediction Error	VaR_{75}
10	1	5226535.8	5415676.6	5733548.4
11	2	4179394.4	4359661.3	4628919.8
12	3	3131667.5	3294105.8	3524291.2
13	4	2127271.9	2249270.9	2419662.6
14	5	1561878.9	1664630.3	1788446.3
15	6	1177743.7	1269130.6	1420236.8
16	7	744287.4	817531.0	946824.5
17	8	445521.3	509135.8	578615.0
18	9	86554.6	113551.8	157804.1
Present Value		17873967	18830614	20255821

Table 7. Present values of the calendar year reserves, of the calendar year reserves plus a 25% of the prediction error and for the Value at Risk with a confidence level of the 75% for the overdispersed Poisson distribution using bootstrap methodology with 1,000 resamples.

Calendar year	Deferral (in years)	Reserve	Reserve + 0.25 Prediction Error	VaR_{75}
10	1	5096855.3	5260096.5	5097605.4
11	2	4050001.5	4186413.5	4050710.2
12	3	3064407.7	3173114.1	3065071.4
13	4	2078010.5	2152405.9	2078475.6
14	5	1510392.7	1568871.3	1510777.7
15	6	1095402.7	1143965.8	1095743.5
16	7	692118.4	727768.8	692406.3
17	8	416539.9	443900.3	416786.1
18	9	82075.9	88738.2	82174.8
Present Value		17310125	17938348	17313858

Table 8. Present values of the calendar year reserves, of the calendar year reserves plus a 25% of the prediction error and for the Value at Risk with a confidence level of the 75% for the Gamma distribution using bootstrap methodology with 1,000 resamples.

4. CONCLUSION

We have deduced the formulas (2) and (3) of the calendar year reserves for the GLM in the general case of the parametric distribution family (1). We have illustrated the analysis with the triangle of Taylor and Ashe (1983). The final results of the calendar years study for the IBNR claim reserving problem gives a tool for the actuary to take decisions about the financial inversions of capital and the solvency margins in the actual context of Solvency II.

ACKNOWLEDGEMENTS

Authors have been supported by the Spanish Ministerio de Educación y Ciencia under grant MTM2010-17323 and by the Generalitat de Catalunya, AGAUR under grant 2014SGR152.

REFERENCES

Boj, E., Claramunt, M. M., and Fortiana, J. (2004). *Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación*. Cuadernos de la Fundación Mapfre, 88. Madrid: Fundación Mapfre Estudios.

Boj, E., and Costa, T. (2014). *Modelo lineal generalizado y cálculo de la provisión técnica*. Depósito digital de la Universidad de Barcelona. Colección de objetos y materiales docentes (OMADO). http://hdl.handle.net/2445/49068

Boj, E., Costa, T., and Espejo, J. (2014). "Provisiones técnicas por años de calendario mediante modelo lineal generalizado. Una aplicación con RExcel". *Anales del Instituto de Actuarios Españoles*, Tercera Época, 20, 83-116.

England, P.D., and Verrall, R.J. (1999). "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". *Insurance: Mathematics and Economics*, 25, 281-293.

England, P.D., and Verrall, R.J. (2002). "Stochastic Claims Reserving in general Insurance" (with discussion). *British Actuarial Journal*, 8, 443-544.

England, P. D. (2002). Addendum to "Analytic and bootstrap estimates of prediction errors in claim reserving". *Insurance: Mathematics and Economics*, 31, 461-466.

England, P.D., and Verrall, R.J. (2006). "Predictive Distributions of Outstanding Liabilities in General Insurance". *Annals of Actuarial Science*, 1 (II), 221–270.

Espejo, J., Boj, E., and Costa, T. (2014). *Una aplicación de RExcel para el cálculo de provisiones técnicas con modelo lineal generalizado*. Depósito digital de la Universidad de Barcelona. Colección de Investigación- Software. http://hdl.handle.net/2445/56230

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models. Second Edition*. London: Chapman and Hall.

Renshaw, A.E. (1989). "Chain Ladder and Interactive Modelling (Claims Reserving and GLIM)". *Journal of the Institute of Actuaries*, 116 (III), 559-587.

Renshaw, A.E. (1994). "On the Second Moment Properties and the Implementation of Certain Glim Based Stochastic Claims Reserving Models". *Actuarial Research Paper*, No. 65. Department of Actuarial Science and Statistics, City University, London.

Taylor, G., and Ashe, F.R. (1983). "Second Moments of Estimates of Outstanding Claims". *Journal of Econometrics*, 23, 37-61.

20. AN APPROACH TO SOLVE THE WEAKNESS OF BLOCK MAXIMA METHOD

A. Cabaña
Universitat Autònoma de Barcelona, Mathematics Department, Spain,
acabana@mat.uab.cat

I. Serra
Centre de Recerca Matemàtica, Barcelona, Spain,
iserra@crm.cat

ABSTRACT

A modification of the standard block maxima modeling procedure for extremes is proposed. It is based on fitting GPD into the blocks and replacing their observed maxima by suitable information on which the likelihood theory for GEV estimation is applied. This approach is developed to obtain the worst possible scenario and applied to analyze a set of rainfall data.

1. INTRODUCTION

When a rare event causes a catastrophe, arises the question "could this has been predicted?". Extreme value theory (EVT) could underestimate the probabilities of extreme events, and this becomes evident when such events do occur and end up in a catastrophe. Coles *et al.* (2003) mention several examples in hydrology, ecology, natural resources management, and insurance.

Classical EVT relies on asymptotic theory for the distribution on the maximum of a series of independent and identically distributed data. It is well known that, if the maximum has a non-degenerate distribution, this distribution belongs to one of

three possible families, Fréchet, Gumbel or Weibull, with very different tail behavior. Moreover, it is not necessary to know the entire distribution of the data in order to determine the asymptotic behavior of the extrema. The tail behavior is determined by a parameter that can be estimated from data. However, if the few available data on extremes decide the value of the shape of the asymptotic distribution, then it can be misleading, see Castillo and Serra (2015). Considering maximum block data dismisses at least half of the data available. On the other hand, inference based on them might suggest lighter tails, and hence underestimate period returns and even extreme. It is natural to think that, depending on the particular situation, the user needs to be more or less demanding: it might be advisable to choose a heavier tail behavior in order to make certain decisions. In case of doubts, the safer decision might be to choose heavier tails than those determined solely by the data. In general, the framework of the data should determine the type of tail, maybe including a bounded support. The standard methodology can be modified in order to take profit of high values other than the maxima, thus admitting a certain control of the "degree of danger" imposed in the model.

2. MODIFIED BLOCK MAXIMA METHODOLOGY

In order to fix notation, we review the well known extreme value methodology. Consider a sample $X1, X2, ..., Xn$ of independent and identically distributed (i.i.d.) random variables (r.v.'s), with unknown common cumulative distribution function (c.d.f.) $F(x) = P(X_i \leq x)$, and define

$$M_n = \max\{X_1, X_2, \ldots, X_n\}$$

The c.d.f. of this random variable is

$$P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = F^n(x) \qquad (1)$$

Even if F were known, the expression for this distribution would be complicated, hence, standard extreme value methods take profit of the asymptotic distribution of F^n. When a c.d.f H and sequences an and $b_n > 0$ exist, such that

$H(x) = \lim_{n\to\infty} P\left(\frac{M_n - a_n}{b_n} \leq x\right) = \lim_{n\to\infty} F^n(b_n x + a_n)$, then Fisher-Tippet Gnedenko's theorem guarantees that H must be of the form

$$H_\xi(z) = \begin{cases} \exp\{-(1+\xi z\})^{-1/\xi} & \xi \neq 0, \quad 1+\xi z > 0 \\ \exp\{-\exp(-z)\} & \xi = 0 \quad -\infty \leq z \leq \infty \end{cases}$$

Parameter ξ is known as the shape, and H is said to belong to the Generalized Extreme Distribution (GEV). When $\xi > 0$ the distribution corresponds to the Frèchet family, with heavy tails, $\xi = 0$ corresponds to the Gumbel family, with exponential tails, and $\xi < 0$ is the Weibull family, with bounded support to the right and called light tails.

The GEV is invariant under scale and location transformations, and so, Fisher-Tippet-Gnedenko's theorem can be interpreted to say that for n large enough, for $x = b_n z + a_n$, $P(M_n x) \approx H_{\xi,a,b}((x - a_n)/b_n) = H_{\xi,a_n,b_n}(x)$, and this expression is used in order to make inference on M_n.

Parameters for the GEV can be estimated by standard methods, and this methodology is useful when only the maximum value in a given block is available. However, when more data in each block are available, they must be taken into account.

It is natural to consider as "extrema" the values X_i that exceed a certain threshold u. The stochastic behavior of the observations over the threshold u are given by the conditional probability $F_u(x) = P(X - u \leq x | x > u) = \frac{F(x+u)-F(u)}{1-F(u)}$, which has a known asymptotic distribution (as $u \longrightarrow \infty$).

When the asymptotic distribution of the maximum M_n is a GEV with shape parameter ξ, the asymptotic distribution (when $u \longrightarrow \infty$) of the observations over the threshold u is a Generalizad Pareto distribution (GPD)(Pickands-Balkema-de Haan's theorem) with the same shape parameter ξ. The cdf of the GPD is

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1+\xi x/\beta)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-x/\beta) & \xi = 0 \end{cases}$$

with $\beta > 0$, and $x \geq 0$ when $\xi \geq 0$, and $0 \leq x \leq -\beta/\xi$ when $\xi < 0$. As in the GEV case, the GPD corresponds to three families of distributions: $\xi > 0$ corresponds to Pareto, $\xi = 0$ is an Exponencial and $\xi < 0$ a bounded support distribution. Further details can be found in Beirlant *et al.* (2004), for instance.

A major weakness of the use of the GEV distribution is that it uses only the maximum and thus many data are wasted. Another problem is the determination of an appropriate block size n, specially in case of time series data where the time dependence is to be thinned out by using appropriate independent blocks from which one extracts one maximum. Further, the selection of blocks must provide the same distribution within blocks in order to apply the classical block maxima methodology. However, there are generalizations on the previous results for non i.d. cases, see Falk *et al.* (2011).

We present in the sequel an alternative approach, modified block maxima (MBM), in order to apply block maxima method for any block size, that also takes profit of other high data, if available. Let $\{X_t\}_t$, $t \in T$ be a sample, for instance from time series. Fix blocks $\{B_s\}_s$ for $s = 1, \cdots, k$. For each block, B_s, several methods provides an estimation of the optimal threshold, $\hat{u}_s$ and a distribution for the tail in the GPD family parameterized by the estimations, $\hat{\xi}_s$ and $\hat{\beta}_s$, $F_s = G_{\hat{\xi}_s, \hat{\beta}_s}$. Then the distribution of the maximum in block s restricted to be bigger than u_s is given by F_s^n, see (1). Moreover, an asymptotic confidence interval of size γ is given by

$$\left(u_s + \hat{\beta}_s/\hat{\xi}_s((1-(1-\gamma)^{1/n_s})^{-\hat{\xi}_s} - 1), u_s + \hat{\beta}_s/\hat{\xi}_s((1-\gamma^{1/n_s})^{-\hat{\xi}_s} - 1)\right).$$

Finally, apply the classical block maximum methodology at the intervals for the maxima rather than the observed maxima.

Notice that the methodologies used in order to search the optimal threshold in each block, to estimate the parameters of the GPD and how to develop the classical block maximum methodology on confidence intervals are not determined here, since there are many adequate choices and no universally accepted solutions. For instance, in next example we consider the methodology in Gomes and Pestana (2007) , maximum likelihood estimation and a framework under the constraint to be a "model for disasters", respectively.

3. DATA ANALYSIS: ANTICIPATING CATASTHOPHES

We show an example on real data of the advantages of the correction proposed in block maximum methodology. In December 15, 1999 the rain changed the shape of the central coast of Venezuela, and caused around 15,000 missing people, and about 3500 millon dollar losses. Alexander von Humboldt mentions a precedent from 1797, and in 1951, have rains in the same area caused a few deaths and other damages, in an area that was barely inhabited by the time.We analyze the record of measurements of the rainfall series recorded at Maiquetia International Airport, on the central coast of Venezuela from 1961 until 1999, see Figure 1. These data have been studied by many authors in order to anticipate catastrophes. Perhaps the most successful approach is that of Coles and Pericchi (2003). They show that the standard GEV approach for the data previous to 1999 lead to completely unbelievable return levels (i.e., fail to capture the 1999 event), and propose a Bayesian model with diffuse prior compatible with the 1999 observed catastrophe.

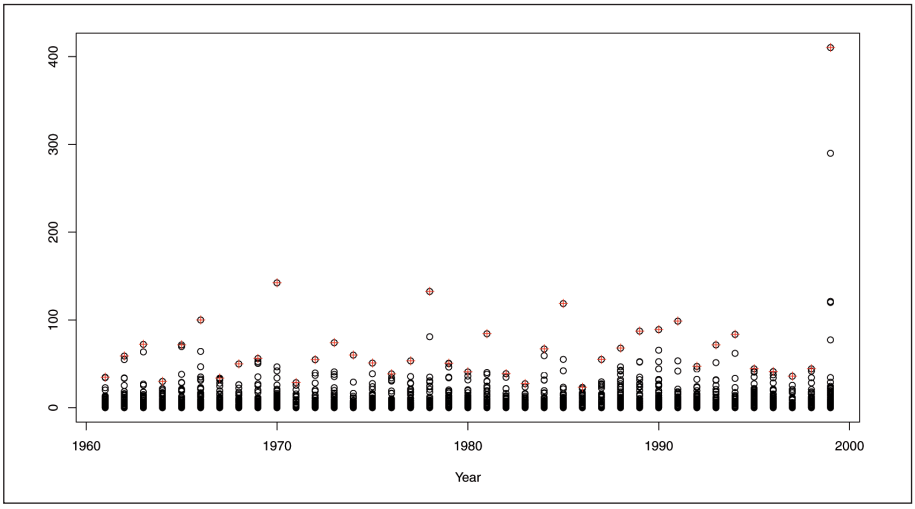


Figure 1. Daily rainfall from 1961 until 1999 on the central coast of Venezuela. The crossed points correspond to annual maxima.

Figure 2 shows that the observation of December 15, 1999 was not likely to have occurred if considering the classical EVT methodology but using our methodology, it

was plausible within only five years. Remark that the MBM was computed under maximum likelihood, see Pawitan (2001), considering the right confidence bound in order to accomplish the framework with the constraint: to be a "model for disasters".

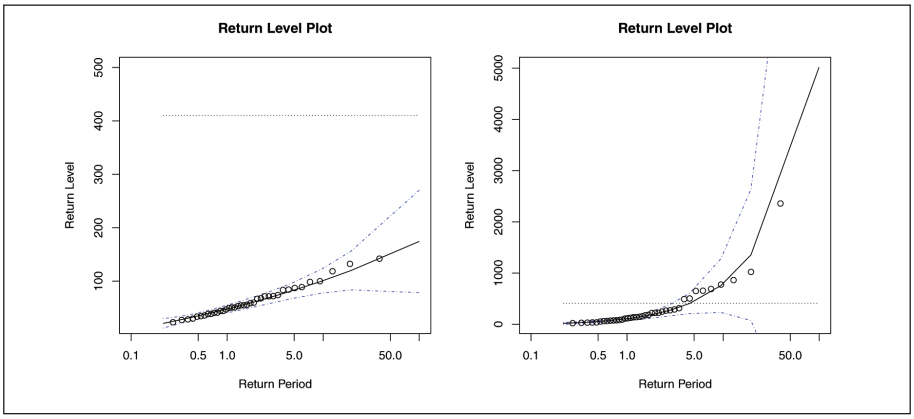


Figure 2. Return level plot excluding the 1999 observed catastrophe, 410mm (showed by horizontal line). Left was computed with classical block maxima methodology and right with MBM.

4. CONCLUSION

As many authors mention, traditional block maxima methodology should be applied only in idyllic scenarios: ensuring i.i.d distribution and non-heavy tails. Alternatively, the following recipe can be applied:

- Choose k blocks.
- Choose a level us $s = 1, \ldots k$ in each block with a suitable methodology.
- Estimate the parameters of the GPD in each block.
- Compute confidence intervals for the maximum.
- Estimate the GEV from suitable information given by the intervals.

This simple procedure makes use of all available data which has influence on maximum. More extensions on MBM can be developed for different frameworks.

ACKNOWLEDGMENTS

Research expenses were founded by projects MTM2012-31118 from Spanish MINECO, 2014SGR1463 from AGAUR and La Caixa foundation.

REFERENCES

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications.* Wiley Series in Probability and Statistics.

Castillo, J. and Serra, I. (2015). "Likelihood Inference for Generalized Pareto Distribution". *Computational Statistics and Data Analysis,* 83, 116-128.

Coles, S., and Pericchi, L.R. (2003). "Anticipating Catastrophes through Extreme Value Modelling". *Journal of the Royal Statistical Society: Series C,* 52, 405-416.

Coles, S., Pericchi, L.R., and Sisson, S. (2003). "A Fully Probabilistic Approach to Extreme Rainfall Modelling". *Journal of Hydrology,* 273, 35-50.

Coles, S. (2001). *An Introduction to Statistical of Extremes Values.* Springer, London.

Falk, M., Hüsler, J., and Reiss, R. (2011). *Laws of Small Numbers: Extremes and Rare Events.* Third, revised and extended edition. Birkhäuser/Springer Basel AG, Basel.

Gomes, M.I., and Pestana, D. (2007). "A Sturdy Reduced-Bias Extreme Quantile (VaR) Estimator". *Journal of the American Statistical Association,* 102, 477.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

21. SOLVING THE MULTI-DEPOT VEHICLE ROUTING PROBLEM CONSIDERING UNCERTAINTY AND RISK FACTORS

L. Calvet, A. Juan and N. Schefers
IN3 - Open University of Catalonia, Computer Science Department, Spain,
{lcalvetl, ajuanp}@uoc.edu, n.schefers@gmail.com

ABSTRACT

This paper describes a hybrid algorithm for solving the Multi-Depot Vehicle Routing Problem with Stochastic Demands. Our approach is based on the Iterated Local Search metaheuristic and Monte Carlo Simulation techniques; while the metaheuristic guides the search to promising solutions, the generation of multiple scenarios allows the estimation of their performance. It assumes a strong correlation between the performance of a solution for the deterministic version of the problem, extensively studied by the scientific community, and for the stochastic counterpart. This assumption enables the usage of already existing efficient heuristics for the deterministic case. To reduce the route-failure risk, safety-stocks are included in our algorithm. It provides the solutions with the lowest expected total costs. A risk analysis is proposed to choose one. A computational experiment is performed to observe the difference of costs when assuming demands with null or positive variances in problem instances characterized by a medium level of uncertainty, and to illustrate the introduction of risk aversion when selecting the best solution.

1. INTRODUCTION

Nowadays, life is plenty of uncertainty. For instance, the punctuality of a flight may be affected by the weather, a portfolio return strongly depends on economical / political events, and the customer demands for the products of a company are highly related to the efforts made by its competitors on marketing strategies. These facts

must be taken into account by decision-makers when addressing the corresponding optimization problems (flight scheduling, capital management, and production planning, among others). Although predictions of the unknown values based on historical data can be computed, even slightly differences between the predicted and the real scenarios may have an important impact on the final output. Moreover, globalization is making systems more complex. As a consequence, there is a growing interest in abandoning the classical approach of assuming that all variables are known, which results in an easier optimization, and adopting a more realistic approach that considers uncertainty (Bianchi *et al.*, 2009).

In this paper, a highly relevant problem of routing is addressed. It is the Multi-Depot Vehicle Routing Problem with stochastic demands (MDVRPSD), which consists in designing a set of routes with an homogeneous fleet of capacitated vehicles that must start and end at the same depot (Figure 1). The aim of this problem is to minimize the total cost of the routes, usually distance-based, while satisfying all customer demands, which are only revealed when a vehicle reaches the corresponding customer place. We propose a simheuristic approach (Juan *et al.*, in press) to address this problem. It consists of an implementation of the Iterated Local Search (ILS) metaheuristic to balance the trade-off between intensification and diversification during the search of a high-quality solution, which includes Monte Carlo Simulation (MCS) techniques to assess the solution performance in a number of generated scenarios.

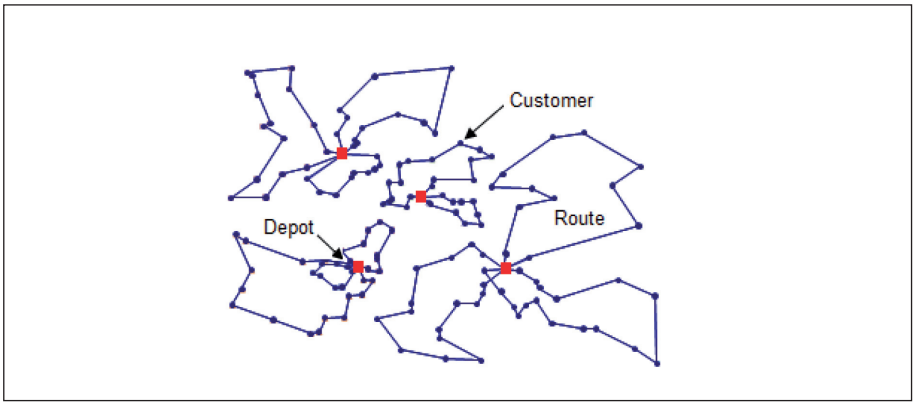


Figure 1. Routing design for a medium-size MDVRPSD instance.

Despite the relevance of the MDVRPSD, few works have been dedicated to its study. For this reason, we highlight also some remarkable proposals for addressing other routing problems considering uncertainty.

Tan *et al.* (2007) present a multiobjective Evolutionary Algorithm that combines heuristics and simulation techniques to solve the Vehicle Routing Problem with Stochastic Demands (VRPSD) with time windows. Moghaddam *et al.* (2012) consider the Particle Swarm Optimization metaheuristic to obtain robust MDVRPSD solutions. In Juan *et al.* (2011), the authors solve the VRPSD by transforming it into Capacitated Vehicle Routing Problems (CVRPs), for which powerful algorithms exist.

2. OUR APPROACH

A two-stage approach is proposed for addressing the MDVRPSD. In the first stage, only the statistical distributions, either empirical or theoretical, of the customer demands are known. A set of routes is designed, supposing that the distribution means are the real demands. The second stage takes place while completing the routes, the demand of each customer is revealed when a vehicle reaches its place. If the remaining vehicle capacity is not enough to satisfy the customer demand (so-called route failure), the vehicle completely unloads, goes back to the depot to re-stock and turns to the customer place to fully satisfy its demand and continue the route as planned. The risk of a route-failure is decreased by considering the option of returning to the depot to re-stock after supplying a customer if the vehicle remaining capacity is lower than the expected demand of the next customer. The actions undertaken to be able to serve all customers minimizing the distance-based cost of the changes applied to the routes planned are corrective actions. Therefore, the first stage provides an aprioristic solution for the MDVRPSD that will be applied as long as there is no need of taking corrective actions. The cost associated to this solution can be considered fixed. Then, the second stage adapts this solution to the specific scenario. The cost of the corrective actions, which is variable, increases the cost.

In order to reduce the risk of having to modify the planned routes, we employ a percentage of the vehicle capacity as a reserve for satisfying demands higher than the

expected, so-called safety stock (Juan *et al.*, 2011). This is done by considering a lower vehicle capacity during the first stage.

Our approach assumes that there is a strong correlation between a solution performance for a MDVRPSD instance considering only the fixed cost (which is the cost when the demand variances are zero) and its performance analyzing the total cost (fixed cost and cost of corrective actions). The first stage is in fact a MDVRP, therefore it makes sense to incorporate efficient heuristics for this problem, much more studied.

The steps of our approach are explained next.

1. Transform the MDVRPSD instance in a MDVRP instance by setting each deterministic demand to the mean of the corresponding statistical distribution.

2. Obtain an initial solution for the MDVRP considering a specific value of safety stock. First, customers are allocated to a depot with a procedure based on the concept of marginal savings (Juan *et al.*, 2014). Then, each resulting CVRP is solved with the CWS heuristic. This initial solution has associated a cost that represents the fixed cost for the initial MDVRPSD instance. The variable cost due to potential corrective actions is estimated with MCS techniques: a low number of scenarios are generated and the average cost is computed.

3. An ILS procedure is started. The best solution and the base solution are set to the initial solution.

   The first step of the iterative process is to perturbate the base solution. It is performed by modifying the assignation of a given percentage of the customers and re-designing the routes. This procedure is implemented as described in (Juan *et al.*, 2014). If the fixed cost of the new solution is lower than the fixed cost of the base solution, the latter is updated. Afterwards, the total cost of the base solution is estimated as before, and the best solution is updated if needed according to the estimated total costs. A demon-based acceptance criterion (Talbi, 2009) is applied in case the new solution is not better than the base solution. The estimated total cost is updated if the base solution is deteriorated. This

process is repeated until a time-based stopping condition is met. The best solution found (or a set of them) is improved by means of a local search. Details of this procedure can be also found in Juan *et al.* (2014). It was designed for the deterministic version of the problem. We have adapted it by storing the best 5 solutions of the local search, estimating the total cost for each one and returning that providing the lowest value.

4. Finally, a high number of simulations is employed to obtain the total cost of the selected solution for a number of scenarios.

3. COMPUTATIONAL EXPERIMENTS

This algorithm has been implemented as a Java application. The data analysis has been performed with R. A standard personal computer, Intel QuadCore i5 CPU at 3.2 GHz and 4 GB RAM with Windows XP, has been used. In order to illustrate our methodology, 6 MDVRPSD instances are solved. They were adapted from 6 MDVRP benchmark instances tested in Juan *et al.* (2014) by considering the demands as the means of Lognormal distributions with a variance equal to 0.5 multiplied by the corresponding mean. The maximum running time of the ILS procedure is set at 60 seconds. 21 equally-spaced values ranging from 0.0 to 0.2 for the percentage of the vehicle capacity reserved for safety stock are tested. The number of simulations are 200 for fast estimations and 2000 for long. 5 seeds are used for solving each instance. Only the best results are stored.

Results are displayed in Table 1. The first two columns correspond to the best solution found considering only fixed costs, i.e. the best solution for the deterministic case. They represent the total costs of the routes under a scenario where variances are null (NV-DC: no variance-deterministic case) or positive (V-DC), respectively. The next two columns (NV-SC and V-SC) contains the same information for the best solution considering estimated total costs, i.e., the stochastic case. The last column shows the difference between the total cost associated to each solution in a scenario with uncertainty in demands. It can be observed that assuming deterministic demands when there is in fact a reasonable level of uncertainty leads to solutions with a worse performance.

Although Table 1 is useful to observe the differences between assuming deterministic demands or not, the decision-maker may use a risk analysis to select the best solution between a set of promising ones. It is important to clearly define the concept of best. The table shows the solutions with the lowest fixed / expected total cost. The multiple box-plot in Figure 2 reveals more information for the instance p01, providing the quartiles and outliers of samples of total costs for the solution with the lowest total cost in the deterministic case, and the five solutions with the lowest expected total costs. The first solution (V-DC) has associated the highest risk, since it provides the minimum and the maximum total costs. In addition, it presents the highest mean and median. The second solution (V-SC1) is the one that has associated the lowest mean, and also the lowest risk. There is a positive correlation between these measures, since there are outliers only with high values. In some cases, there is a trade-off between mean and risk. For instance, while the fifth solution (V-SC4) has a lower mean than the sixth (V-SC5), the risk is indeed bigger. Therefore, it may be needed to take into account the risk aversion of the decision-maker to choose the final solution.

Instance	NV-DC	V-DC (1)	NV-SC	V-SC (2)	Gap % (2-1)
p01	588.5	624.5	601.0	610.3	-2.3
p02	481.2	505.2		484.4	-4.1
p03	648.8	673.7	650.5	659.2	-2.2
p04	1045.4	1193.8	1060.9	1135.9	-4.9
p05	777.6	815.9	772.	3 796.2	-2.4
p06	897.6	994.9	927.1	947.6	-4.8

Table 1. Total costs for different scenarios and instances.

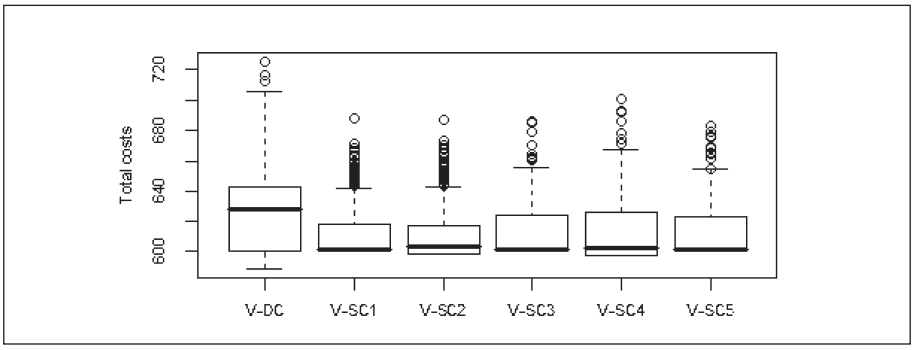


Figure 2. Risk analysis of alternative solutions.

4. CONCLUSIONS

This paper has presented a hybrid algorithm for solving the Multi-Depot Vehicle Routing Problem considering uncertainty and risk issues. Our proposal combines the Iterated Local Search metaheuristic with Monte Carlo Simulation (MCS) techniques. The experimental results show that even in a scenario characterized by demands with moderated variances, assuming that demands are deterministic may result in a solution with a poor performance, considering the expected total cost and the risk level associated. Since our algorithm forces the final solution to provide the best expected performance for a number of scenarios, robust solutions are returned. Regarding the future work, it is desirable to continue addressing optimization problems under more realistic assumptions, and introducing the decision-maker willingness to accept risk.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (grant TRA2013-48180-C3-3-P) and FEDER. Likewise we want to acknowledge the support received by the Department of Universities, Research & Information Society of the Catalan Government (Grant 2014-CTP-00001).

REFERENCES

Bianchi, L., Marco, D., Gambardella, L.M., and Gutjahr, W.J. (2009). "A Survey on Metaheuristics for Stochastic Combinatorial Optimization". *Natural Computing*, 8, 239-287.

Juan, A., Faulin, J., Grasman, S., Rabe, M., and Figueira, G. (in press). "A Review of Simheuristics: Extending Metaheuristics to Deal with Stochastic Combinatorial Optimization Problems". *Operations Research Perspectives*.

Juan, A., Faulin, J., Grasman, S., Riera, D., Marull, J., and Mendez, C. (2011). "Using Safety Stocks and Simulation to Solve the Vehicle Routing Problem with Stochastic Demands". *Transportation Research Part C*, 19, 751-765.

Juan, A., Pascual, I., Guimarans, D., and Barrios, B. (2014). "Combining Biased Randomization with Iterated Local Search for Solving the Multi-Depot Vehicle Routing Problem". *International Transactions in Operational Research.* DOI: 10.1111/itor.12101.

Moghaddam, B. F., Ruiz, R., and Sadjadi, S. J. (2012). "Vehicle Routing Problem with Uncertain Demands: An Advanced Particle Swarm Algorithm". *Computers & Industrial Engineering*, 62, 306-317.

Talbi, E. (2009). *Metaheuristics: From Design to Implementation.* New Jersey: John Wiley & Sons.

Tan, K.C., Cheong, C.Y., and Goh, C.K. (2007). "Solving Multiobjective Vehicle Routing Problem with Stochastic Demand via Evolutionary Computation". *Discrete optimization,* 177, 813-839.

22. FIRST HITTING TIME REGRESSION FOR RECURRENT EVENTS WITH RANDOM EFFECTS

C. Caroni
Department of Mathematics, National Technical University of Athens, Greece, ccar@math.ntua.gr

S. Malefaki
Department of Mechanical Engineering and Aeronautics, University of Patras, Greece, smalefaki@upatras.gr

P. Economou
Department of Civil Engineering, University of Patras, Greece, peconom@upatras.gr

ABSTRACT

The model proposed here for analysing lifetime data with recurrent events extends a first hitting time regression model with individual random effects. Successive event times are modeled as realizations of the first hitting times of an underlying Wiener process, leading to Inverse Gaussian distributions of times between events. The parameters of the distribution may depend on features of the process (such as number of previous events and total elapsed time) as well as on measured covariates, which may be time varying, and the individual's random effects. A Bayesian approach is adopted for model estimation using an MCMC algorithm. The model is applied to a database on drug users who made repeated contacts with treatment services.

1. INTRODUCTION

Lifetimes can sometimes be represented as the outcomes of an underlying stochastic process (Aalen and Gjessing, 2001; Lee and Whitmore, 2006). For example, if a subject dies when his or her latent health status falls to zero, the lifetime is the first hitting time on the boundary at zero of the process which describes health status. We analyse here a regression model in which the underlying stochastic process is a Wiener process. Lifetimes therefore follow an Inverse Gaussian distribution which depends on two parameters, corresponding to the drift and starting level of this process. Both of the parameters are allowed to depend on covariates through linear predictors and appropriate link functions (Lee *et al.*, 2004; Lee and Whitmore, 2006). This model has been extended in various directions, including the introduction of random effects in order to allow additional heterogeneity. Pennell *et al.* (2010) presented a model with individual random effects in both the drift and starting level. We propose a model that generalizes this in order to model recurrent events such as repeated periods of illness or the breakdowns of a repairable system. The analysis is performed within a Bayesian framework using an MCMC algorithm for estimation.

2. THRESHOLD REGRESSION MODEL WITH RANDOM EFFECTS

We suppose that the data for the i^{th} of n individuals consist of n_i event times $0 < t_{1i} < t_{2i} < \ldots < t_{ni\ i}$, with indicator $d_i = 0$ (censored) or 1 (failed) for the final event. These are realizations of the underlying stochastic process. The covariate vector z_{mi} corresponds to the m^{th} stage of the i^{th} individual.

The m^{th} stage of the process is a Wiener process with initial state x_m, variance σ^2 and drift μm. If $\mu m \leq 0$, it will eventually reach the threshold at zero with probability 1 and the time $S_m = T_m - T_{m-1}$ when this first happens follows an Inverse Gaussian distribution with mean $-\frac{x_m}{\mu_m}$ and scale parameter $\left(\frac{x_m}{\sigma}\right)^2$. If $\mu m > 0$, there is a non-zero probability equal to $1-\exp(-2x_m\ \mu_m/\sigma^2)$ that the m^{th} stage of the Wiener process never reaches zero. However, the same Inverse Gaussian distribution applies conditionally on reaching the boundary. Because the scale of the underlying latent process is arbitrary, one parameter can be removed by fixing the variance, $\sigma^2 = 1$, say.

We allow the location parameters of both the starting level and the drift to depend on covariates u_{mi} and v_{mi}i, which may change from stage to stage, as well as on the number of previous stages and the total time on observation until the beginning of each stage. The sets u_{mi} and v_{mi} are subsets of the available covariates z_{mi}, not necessarily disjoint.

We assume a linear regression model for the drift parameter

$$\mu_{mi} = \sum_{k=1}^{m} b_{ki} + \tilde{\mathbf{v}}'_{\mathbf{mi}}\boldsymbol{\beta}^*$$

for every m $(1 \leq m \leq ni)$, where

$$\tilde{\mathbf{v}}'_{\mathbf{mi}}\boldsymbol{\beta}^* = \mathbf{v}'_{\mathbf{1i}}\boldsymbol{\beta} + (\mathbf{v}'_{\mathbf{mi}} - \mathbf{v}'_{\mathbf{1i}})\boldsymbol{\beta}_{\boldsymbol{d}} + \beta_{ns}(m-1) + \beta_s t_{(m-1)i}\,.$$

The coefficients with subscripts ns and s introduce the dependence on the number of previous stages and the time elapsed, respectively, while the coefficients with subscripts d express the influence of the difference between the values of the covariates of the m_{th} stage and the initial stage. We model the random starting level as

$$x_{mi} \sim N_+\left(\tilde{\mathbf{u}}'_{\mathbf{mi}}\boldsymbol{\alpha}^*, \tau^{-1}\right)$$

where N+() denotes a normal distribution truncated to the left at zero (because xmi is required to be positive) and

$$\tilde{\mathbf{u}}'_{\mathbf{mi}}\boldsymbol{\alpha}^* = \mathbf{u}'_{\mathbf{1i}}\boldsymbol{\alpha} + (\mathbf{u}'_{\mathbf{mi}} - \mathbf{u}'_{\mathbf{1i}})\boldsymbol{\alpha}_{\boldsymbol{d}} + \alpha_{ns}(m-1) + \alpha_s t_{(m-1)i}\,.$$

When covariates are recorded only at the beginning of the study, as in the application that follows later, the model simplifies because $v_{mi} - v_{1i} = 0$ and $u_{mi} - u_{1i} = 0$. Various further simplifications are possible.

3. INFERENCE

Every stage is described by a Wiener process which given its starting point and drift is independent of other stages. The likelihood is

$$\prod_{i=1}^{n}\left\{\left[\prod_{m=1}^{n_i-1}\frac{x_{mi}}{\sqrt{2\pi\sigma^2 s_{mi}^3}}\exp\left(-\frac{(\mu_{mi}s_{mi}+x_{mi})^2}{2\sigma^2 s_{mi}}\right)\right]\cdot\right.$$
$$\left(\frac{x_{n_i i}}{\sqrt{2\pi\sigma^2 s_{n_i i}^3}}\exp\left(-\frac{(\mu_{n_i i}s_{n_i i}+x_{n_i i})^2}{2\sigma^2 s_{n_i i}}\right)\right)^{d_i}\cdot \tag{1}$$
$$\left.\left(\frac{1}{\sqrt{2\pi\sigma^2 s_{n_i i}}}\exp\left(-\frac{(x_i-x_{n_i i}-\mu_{n_i i}s_{n_i i})^2}{2\sigma^2 s_{n_i i}}\right)\left[1-\exp\left(-\frac{2x_i x_{n_i i}}{\sigma^2 s_{n_i i}}\right)\right]\right)^{1-d_i}\right.$$

The first term in the product is the Inverse Gaussian pdf, contributed by the first $n_i - 1$ uncensored stages. (A more general model can be written down which allows censoring in any stage). The last stage, if uncensored, contributes the second term, or the third term if censored. Under a frequentist approach, the latter would be the survival function of the Inverse Gaussian distribution. What has been done here instead is to treat the unobserved level xi for the censored last stage as an unknown parameter (see Pennell *et al.*, 2010).

Conjugate priors are assigned to the regression parameters $\beta^{*\prime}$ and α^*:

$$\boldsymbol{\beta}^* \sim N_q(\boldsymbol{\beta}^*_{\boldsymbol{prior}}, \boldsymbol{\Sigma}_{\beta^*})$$
$$\boldsymbol{\alpha}^* \sim N_p(\boldsymbol{\alpha}^*_{\boldsymbol{prior}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}^*})\,.$$

Lack of prior knowledge could reasonably be represented by taking $\Sigma\beta^* = c\mathrm{I}q$, where c is a large number such as 104, with a similar choice for α^*. In the absence of information available from previous studies or elsewhere, a diffuse form of the conjugate prior such as Gamma(2, 1/2) or Gamma(2, 1) can be used for λ. Instead of also assigning a Gamma prior to the parameter τ, we follow Pennell *et al.* (2010)'s recommendation to avoid identifiability problems by taking a fixed moderately small value, which they justified on the grounds that results are not sensitive to the choice of τ. Finally, for the unobserved levels x_i we assign an improper, non-informative flat prior to x_i, given $d_i = 0$:

$$\pi(x_i|d_i = 0, \cdot) \propto 1(x_i > 0) \tag{2}$$

since no information is expected to be available for the unobserved levels xi of the stochastic process of the last censored stage. Here $1(\cdot)$ stands for the indicator function.

A hybrid Gibbs Sampler is used for simulation from the full posterior likelihood because some of the full conditional distributions are not available in closed form. The level x_i at the censoring time for the last censored state is updated using a modification of Pennell *et al.*'s (2010) rejection sampling algorithm. A similar approach is used to update the starting level x_{mi}. The regression coefficients for starting levels α^* cannot be sampled from their full conditional posterior by a Gibbs Sampler because of normalizing constants in the denominators of the truncated normal pdfs of x_{mi}. This obstacle can be overcome by using an auxiliary variable approach proposed by Griffiths (2002). The regression coefficients associated with the drifts, the b_{mi} , $1 \leq m \leq n_i$, $1 \leq i \leq n$ and λ are simulated using their full conditional distributions, which are Normal and Gamma, respectively.

4. APPLICATION

The model is illustrated using data on applications for treatment by 1497 drug users in Greece from 2001 onwards. The event of interest is making a new application for treatment. The greatest number of events recorded for one person was eight. The underlying latent stochastic process can be interpreted as the psychological and physical status of the drug user, and the new application for treatment occurs when the process has fallen to zero. For the i^{th} individual, time starts at his or her first entry to the database. The first event time t_{1i} is the time elapsed from then until the next application, if there is one, and so on. Data collection ended in 2011 at which point all times since the last event were censored.

Five covariates were recorded at the beginning of the study: place of residence (Location: 1 = Athens/Piraeus, 2 = Thessaloniki, 3 = Other); gender (1 = male, 0 = female); age in years; the primary substance of abuse of the drug user (PS: 1 = heroin, 0 = other); and year of first entry in the database, recoded by subtracting 2001. The location was represented by two dummy variables: L1 (1=Greater Athens, 0=Other) and L2 (1=Thessaloniki, 0=Other). In addition to these we also included as covariates the number of previous stages (applications for treatment) and the total

time since entering the database. Therefore the covariate vector of an individual's m^{th} stage is

$$(1, L1, L2, Gender, Age, PS, Years, m-1, t_{m-1}).$$

Every covariate was allowed to influence both the starting level and the drift of the process. Consequently, the model contains 19 parameters: 9 for the drift, 9 for the initial stage and one for λ, the inverse of the variance of b_{mi}. We assigned flat priors to the regression parameters with

$$\begin{aligned} \alpha^*_{prior} &= (1,0,0,0,0,0,0,0,0), \ \ \mathbf{\Sigma}_{\boldsymbol{\alpha}^*} = 10^4\mathbf{I}_9 \\ \beta^*_{prior} &= (1,0,0,0,0,0,0,0,0), \ \ \mathbf{\Sigma}_{\boldsymbol{\beta}^*} = 10^4\mathbf{I}_9 \\ \lambda_{prior} &\sim Gamma(2, 1/2) \end{aligned}$$

θ	Mean	Mode	Q_1	Median	Q_3	P($\theta < 0$|Data)
$\alpha 0$	21.711	21.948	19.925	21.660	23.384	←0.0001
$\alpha 1$	-3.508	-3.853	-4.621	-3.520	-2.225	0.984
$\alpha 2$	-0.311	-0.619	-1.420	-0.303	0.947	0.566
$\alpha 3$	1.086	1.114	0.627	1.097	1.555	0.059
$\alpha 4$	-0.173	-0.170	-0.193	-0.172	-0.152	→0.9999
$\alpha 5$	-5.338	-5.189	-6.271	-5.276	-4.358	0.999
$\alpha 6$	1.288	1.307	1.140	1.289	1.424	←0.0001
$\alpha 7$	1.670	1.708	1.246	1.669	2.074	0.001
$\alpha 8$	0.002	0.002	0.001	0.002	0.003	0.057
$\beta 0$	0.034	0.031	0.019	0.034	0.050	0.070
$\beta 1$	0.032	0.031	0.023	0.032	0.041	0.007
$\beta 2$	-0.023	-0.022	-0.031	-0.023	-0.014	0.962
$\beta 3$	-0.003	-0.004	-0.009	-0.003	0.004	0.605
$\beta 4$	$9.1\cdot 10^{-4}$	$9.0\cdot 10^{-4}$	$6.1\cdot 10^{-4}$	$9.0\cdot 10^{-4}$	0.001	0.015
$\beta 5$	-0.010	-0.008	-0.018	-0.009	-0.002	0.791
$\beta 6$	0.010	0.010	0.009	0.010	0.011	←0.0001
$\beta 7$	0.058	0.057	0.053	0.058	0.063	←0.0001
$\beta 8$	$9.7\cdot 10^{-6}$	$9.4\cdot 10^{-6}$	$2.9\cdot 10^{-6}$	$9.6\cdot 10^{-6}$	$1.64\cdot 10^{-5}$	0.168
λ	150.769	148.093	141.72	150.31	159.385	

Table 1. Descriptive statistics for the posterior distributions of the parameters of the model fitted to the data on drug users.

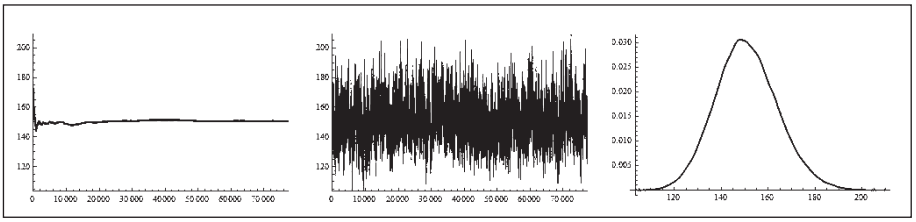


Figure 1. The ergodic mean (left plot), the trace (middle plot) and a smooth kernel estimate of the density of the posterior distribution (right plot) of λ of the model fitted to the data on drug users.

and we ran the MCMC for 90,000 iterations, the first 15,000 of which were discarded as burn-in.

Table 1 presents descriptive statistics of the posterior distribution of the parameters. Figure 1 presents the ergodic mean (left plot), the trace (middle plot) and a smooth kernel estimate of the density of the posterior distribution (right plot) for the parameter λ, as an illustration of results. Plots for other parameters are similar.

From the values of $P(\theta < 0|\text{Data})$, it can be seen which covariates appear to affect event times. There is no strong evidence of any effect of L2, gender and total time on study t_{m-1}. L1, age, PS, years and number of events $m-1$ appear to affect one or both parameters of the underlying process. For example, if the primary substance is heroin (PS = 1), the starting level appears to be lower, which is likely to lead to shorter times between events compared to users of other substances, other things being equal. This covariate does not appear to affect the drift of the process. The number of events appears to be associated with both a higher starting level and a higher value of the drift (i.e. a more positive value, taking the process away from zero); both imply that users with a longer history will have longer intervals between events. Because the estimate of λ is large, the variance of the random effects is close to zero. This suggests that it might not be necessary to include individual random effects b_{mi} in order to model these data.

5. CONCLUSION

In first hitting time regression models, times-to-event are modeled as realizations of a latent process that corresponds to a real feature of the actual situation, such as

an individual's state of health or, in the application given here, need for treatment. This offers an attractive alternative to lifetime data analysis based on concepts such as hazard rates. It also offers an alternative to the common assumption of proportional hazards. The version of the model proposed here is a flexible model for describing complex data sets containing times to a recurrent event. As well as allowing lifetimes to depend on measured covariates, it also includes individual random effects to allow for additional heterogeneity. As with other regression models for lifetime data, this dependence is introduced through the parameters of the lifetime distribution. In addition, the parameters are allowed to depend on the process (number of previous events, total elapsed time, etc.). The covariates may take different values for different times-to-event of the same individual. Further work is needed on various aspects of these models, including the development of diagnostic tools and model selection procedures.

REFERENCES

Aalen O.O., and Gjessing H.K. (2001). "Understanding the Shape of the Hazard Rate: a Process Point of View". *Statistical Science,* 16, 1-14.

Griffiths W. (2002). "A Gibbs Sampler for the Parameters of a Truncated Multivariate Normal Distribution". *Department of Economics-Working Papers Series* 856, University of Melbourne.

Lee M-L.T., Garshick E., Whitmore G.A., Laden F., and Hart J. (2004). "Assessing lung cancer risk to rail workers using a first hitting time regression model". *Environmetrics,* 15, 501-512.

Lee M-L.T., and Whitmore G.A. (2006). "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary". *Statistical Science,* 21, 501-513.

Pennell M.L., Whitmore G.A., and Lee M-L.T. (2010). "Bayesian Random-Effects Threshold Regression with Application to Survival Data with Non-Proportional Hazards". *Biostatistics,* 11, 111-126.

23. NON-STATIONARY MODELLING OF EXTREMES OF PRECIPITATION AND TEMPERATURE OVER MOUNTAINOUS AREAS UNDER CLIMATE CHANGE

C. Caroni
Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Greece, ccarl@math.ntua.gr

D. Panagoulia
Department of Water Resources and Environmental Engineering, School of Civil Engineering,National Technical University of Athens, Greece, dpanagl@hydro.civil.ntua.gr

P. Economou
Department of Civil Engineering, School of Engineering, University of Patras, Greece, peconom@upatras.gr

ABSTRACT

The generalized extreme value (GEV) distribution is often fitted to environmental time series of extreme values such as annual maxima and minima of temperatures or rainfall. It is often necessary to allow the distribution's parameters to depend on time or other covariates. Important methodological issues include the selection of the best among several models that allow non-stationary of various degrees in the scale and location parameters. The results of simulation studies showed that both the corrected AIC and BIC always detected non-stationarity but the BIC selected the correct model more often except in very small samples. Furthermore, accurate methods of constructing confidence intervals (CIs) for model parameters and other quantities such as the return levels are required. Because normal approximations may be inadequate, bootstrap methods are recommended. Simulation studies indicate that the bias corrected and accelerated (BCa) method is best overall for the

extreme percentiles that are often the focus of interest. A case study is presented of temperature and rainfall extremes over mountainous areas of Greece. Model fitting supports a non-stationary GEV model for temperature with the location parameter depending linearly on time, showing an increasing trend in annual maxima and declining trend in annual minima. On the other hand, the rainfall series appeared to be stationary, reducing to the Gumbel distribution (GEV with shape parameter equal to zero). Fitting within the framework of GAMLSS models showed similar results for several other distributions.

1. INTRODUCTION

The study of extremes in hydro-meteorological series, such as annual or monthly maxima of precipitation or temperature, is fundamental for engineering design. Often, these series are not stationary, because of natural climate variability or anthropogenic climate change (Jain and Lall, 2001; Milly *et al.*, 2008). Consequently, the assumption of a series of independently and identically distributed values over time (stationarity) may need to be modified. Introducing non-stationarity within the framework of standard statistical distributions requires extended models with covariate-dependent changes in one or more of a distribution's parameters (Coles, 2001).

2. STATISTICAL MODELLING

2.1. Generalized extreme value (GEV) distribution

The GEV distribution is widely employed for modelling extremes in the environmental sciences and elsewhere (Reiss and Thomas, 2007). It depends on three parameters: location μ , scale σ and shape ξ . In the non-stationary GEV distribution (El Adlouni *et al.,* 2007; Leclerc and Ouarda, 2007), these are expressed as a function of time t and possibly other covariates (Coles, 2001). We shall only consider dependence on time. Furthermore, we will allow non-stationarity of the location and scale parameters, but not of the shape parameter. This non-stationary GEV $(\mu(t),\sigma(t),\xi)$ distribution has distribution function

$$F\left(y;\mu(t),\sigma(t),\xi\right)=\exp\left\{-\left[1+\xi\frac{y-\mu(t)}{\sigma(t)}\right]^{-1/\xi}\right\}.$$

Specifically, the following regression structures will be considered for the location and scale parameters

$$\mu(t)=\mu_0+\mu_1 t+\mu_2 t^2+\mu_3 t^3$$
$$\ln\sigma(t)=\sigma_0+\sigma_1 t+\sigma_2 t^2+\sigma_3 t^3$$

allowing up to cubic dependence on time t. We denote by GEVjk the model with time dependence of order j in the location parameter and order k in the scale parameter. A convenient tool for fitting stationary and non-stationary GEV distributions by maximum likelihood is the *gev.fit routine* in the R package 'ismev' (http://cran.R-project.org/package=ismev).

2.2. GAMLSS

The class of generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005) provides a flexible framework for modelling, allowing up to four parameters of a distribution chosen from a very wide family to depend on covariates. For example, Villarini *et al.* (2009) examined the fit of five distributions (Gumbel, Weibull, Gamma, Logistic and Lognormal) to data on rainfall and temperature in Rome. Non-stationary GEV distributions can also be fitted within this framework.

The general format of the model for parameter θ_k is

$$g_k\left(\theta_k\right)=X_k\beta_k+\sum_{j=1}^{J_k}Z_{jk}\gamma_{jk}$$

where g_k is a link function, X_k is a design matrix containing the values of J_k covariates for each of n independent observations, β_k is a parameter vector of length J_k, Z_{jk} is another known design matrix of dimension $n \times q_{jk}$ and γ_{jk} is a q_{jk} - dimensional random vector.

GAMLSS modelling is implemented in the R package 'gamlss' (http://cran.R-project.org/package=gamlss), which makes it easy to include features such as random effects or non-polynomial dependence on covariates by means of splines.

2.3. Model selection

It is important to have objective procedures for selecting the best fitting among various candidate models. The likelihood ratio test can be used to compare the goodness-of-fit of two hierarchically nested models. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also widely applied for model selection. If $\hat{\ell}$ is the maximized value of the likelihood from a model that contains p parameters, and n is the sample size, these criteria are defined as

$$AICc = -2\hat{\ell} + 2p + \frac{2p\ (p+1)}{n-p-1}$$

(this is the corrected AIC – the third term is a small-sample adjustment) and

$$BIC = -2\hat{\ell} + p\log n\,.$$

The preferred model minimizes the chosen criterion although attention should also be paid to models with values close to the minimum. More details of model selection procedures can be found in Claeskens and Hjort (2008), for example.

Panagoulia *et al.* (2014) carried out a simulation study in order to evaluate empirically the performance of the AICc and BIC in identifying the true model among the set of models GEV*jk* (*j*=0, 1, 2, 3; *k*=0, 1, 2, 3), for samples of sizes *n* = 20, 50 or 100. Both criteria had high success rates in detecting non-stationarity. The BIC was the more successful in identifying the correct model: over 80% of the time for *n*=50 and over 90% for *n*=100. Neither performed well for *n*=20, which is a very small sample in relation to the number of parameters in some of these models.

2.4. Confidence intervals

One of the major objectives of fitting a distribution to climate data is to obtain estimates of its quantiles, especially those related to the return periods of extreme

events: for example, the upper 1% point of the distribution of annual maxima corresponds to a 1/.01 = 100-year return period. Because parametric confidence intervals based on a normal distribution approximation are not expected to work well for extreme quantiles, their construction by bootstrap methods has been examined for GEV models, first by Kysely (2008) in the stationary case and subsequently by Panagoulia *et al.* (2014) in the non-stationary case. Amongst several methods compared, the best was found to be the parametric bootstrap with confidence intervals constructed by the bias corrected and accelerated (BCa) technique.

3. CASE STUDY

Time series of meteorological data are available from several locations within a mountainous area of Greece. Analyses are carried out for annual maxima and minima of temperature and maxima of rainfall. We describe briefly here the results obtained from modeling the annual extremes of temperature in the entire area over 21 years. The series do not appear to be stationary, as the GEV10 model offers significantly improved fit over the GEV00 model (p=0.05 for maximum temperatures, p=0.01 for minima). The smoother fitted to the annual minima in Figure 1 demonstrates the decreasing trend. The suggestion of greater variance in the later years is not borne out by statistical tests (p=0.40 for GEV11 versus GEV10). In contrast to the results of the analysis of rainfall data in Panagoulia *et al.* (2014), the GEV model for temperatures did not reduce to the Gumbel ($\xi = 0$) as the former had significantly better fit than the latter (AIC 96.8 compared to 102.1).

GAMLSS modelling showed that the Inverse Gaussian, Gamma and Lognormal distributions with linear trend in their location parameters all fitted equally well as the GEV10. The introduction of linear dependence on time in the scale parameter did not improve fit.

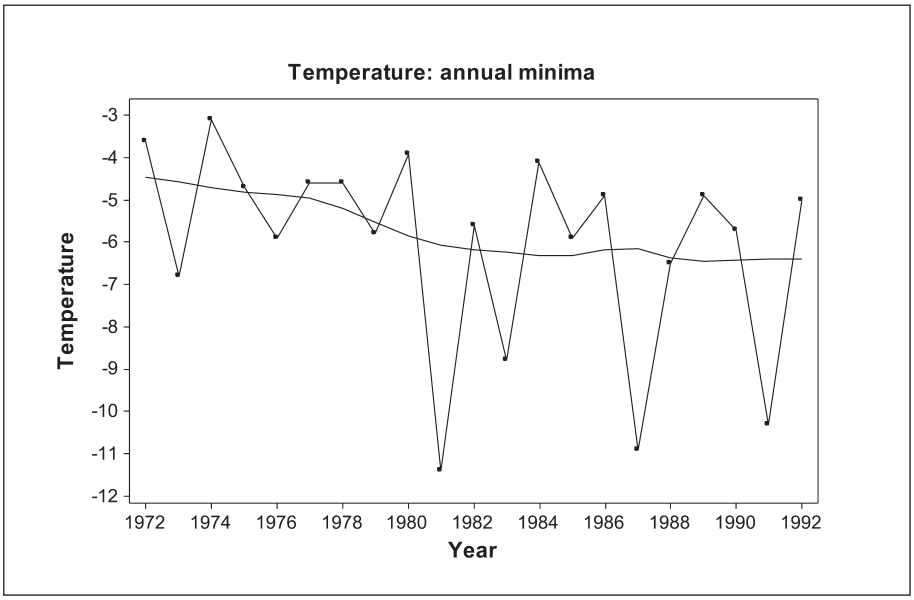


Figure 1. Annual minimum temperatures (ºC) over the whole study area, with trend fitted by locally weighted scatterplot smoothing (lowess: Minitab 14).

REFERENCES

Claeskens, G., and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Coles, G.S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. New York: Springer.

Jain, S., and Lall, U. (2001). "Floods in a Changing Climate: Does the Past Represent the Future?". *Water Resources Research*, 37, 3193-3205.

Kysely, J. (2008). "A Cautionary Note on the Use of Nonparametric Bootstrap for Estimating Uncertainties in Extreme-Value Models". *Journal of Applied Meteorology and Climatology*, 47, 3236-3251.

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, M., Kundzewicz, Z.W., Lettenmaier, D.P., and Stouffer, R.J. (2008). "Stationarity is Dead: Whither Water Management?". *Science*, 319, 573-574.

Panagoulia, D., Economou, P., and Caroni, C. (2014). "Stationary and Non-Stationary GEV Modelling of Extreme Precipitation over a Mountainous Area under Climate Change". *Environmetrics*, 25, 29-43.

Reiss, R.D., and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields* (3rd ed.). Basel: Birkhauser.

Rigby, R.A., and Stasinopoulos, D.M. (2005). "Generalized Additive Models for Location Scale and Shape". *Applied Statistics*, 54, 507-554.

Villarini, G., Smith, J.A., Serinaldi, F., Bales, J., Bates, P.D., and Krajewski, W.F. (2009). "Flood Frequency Analysis for Non-Stationary Annual Peak Records in an Urban Drainage Basin". Advances in Water Resources, 32, 1255-1266.

24. MODELING TRENDS AND INEQUALITY OF LONGEVITY IN THE EUROPEAN UNION COUNTRIES

P. Carracedo and A. Debón
Universitat Politècnica de València, Centro de Gestión de la Calidad y del Cambio, Spain,
patcarga@posgrado.upv.es, andeau@eio.upv.es

S. Haberman
City University London, Cass Business School, United Kingdom,
S.Haberman@city.ac.uk

ABSTRACT

Comparisons of differential survival by country are useful in many domains. In the area of public policy, they help policymakers and analysts assess how much various groups benefit from public programs, such as Social Security and health care. In financial markets and especially for actuaries, they are important for designing annuities and life insurance. We present a method for clustering information about differential survival by country. Then this approach is used to group mortality surfaces for European Union countries. This paper reviews mortality indicators to study mortality inequalities and trends in European Union countries. Additionally, the indicators permit us to characterization of each group. All these statistical analysis were performed using the R environment for statistical computing.

1. INTRODUCTION

Interest in health inequalities between European Union (EU) countries and their regions as well as the various social clusters in the EU population is growing (Spinakis *et al.*, 2011). The problem of determining which mortality indicators influence the

characterization of a country group has been considered from different perspectives and statistical methodologies. The investigation of this issue has made frequent use of traditional statistical methods such as "Classification And Regression Trees" (CART). Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of CART at training time and out-putting the class that is the mode of the classes output by individual trees Liaw and Wiener (2002).

Working in the framework of multi-population mortality models, Hatzopoulos and Haberman (2013) present a new common mortality modeling structure to analyse mortality dynamics for a pool of countries. However, there is a research gap in the characterization of country groups based on mortality indicators and their capacity for discrimination. For this reason, this study also introduces classification and regression trees (CART) and random forests, which can help by allowing us to describe groups and select indicators.

2. ANALYSIS OF MORTALITY DATA FROM THE EUROPEAN UNION

In this section, a method for grouping countries from the distances between their mortality surfaces is presented. A mortality surface is a function of the (time t, age x) "Lexis plane" which describes the mortality experience of a population during a given interval of time. Examples are: the logarithm of the force of mortality, $log(\mu_{xt})$, and the logit transformation of the specific probability of death, $log(q_{xt}/(1-q_{xt}))$. A mortality surface is drawn when the logarithm of the forces of mortality, or the logit transform of the specific probability of death, are plotted against age and time period simultaneously in a three-dimensional diagram. In this paper we are going to show the methodology by using $log(q_{xt}/(1-q_{xt}))$. Additionally, cluster validity indices are introduced, which assess the average compactness and separation of fuzzy partitions generated by the fuzzy c-means algorithm.

CART or trees are going to use for the characterization of groups. Tree models are quite popular as the structure is easier for non technical people to understand and they are implemented in R-packages. Recently there has been a lot of interest in "ensemble learning" Ů methods that generate many classifiers and aggregate their

results (Liaw and Wiener, 2002) called random forest. In standard trees, each node is split using the best split between all the variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. In addition, it is very user-friendly as it only has two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values. The randomForest package provides an R interface to the Fortran programs by Breiman and Cutler (available at http:// www.stat.berkeley.edu/ users/breiman/). (Liaw and Wiener, 2002) provide a brief introduction to the use and features of the R functions. The randomForest package optionally produces two additional pieces of information: a measure of the importance of the predictor variables, and a measure of the internal structure of the data (the proximity of different data points to one another). The reader is referred to (Liaw and Wiener, 2002) for their definitions.

These statistical tools are going to be applied to the most usual mortality indicators such as life expectancy, modal age at death and Gini Index can be calculated from a dynamic period life table, details in Debón *et al.* (2012).

2.1. Data

All these methods will be used to study mortality for the countries of EU for men. The application is carried out using the language and environment for statistical computing and graphics R. We use mortality data from the Human mortality database. There are no complete data for all the EU countries and available countries have a common time range 1990-2009 for the age ranges from 0 to 110 years old.

2.2. Clustering of the countries of EU using Principal Components

We want to gather the 21 countries of the dataset into a number of clusters which would correspond to different mortality profiles. Cluster analysis is a reasonable approach to separate the EU countries into clusters with similar mortality surface summarized in the corresponding 5 CPs.

	1	2	3
Austria	0.01	0.99	0.00
Belgium	0.03	0.97	0.01
Bulgaria	0.66	0.09	0.25
Czech Republic	0.80	0.12	0.08
Denmark	0.06	0.92	0.02
Estonia	0.15	0.05	0.81
Finland	0.11	0.86	0.03
France	0.11	0.84	0.05
Germany	0.03	0.96	0.01
Hungary	0.73	0.08	0.19
Ireland	0.21	0.72	0.07
Italy	0.05	0.93	0.02
Latvia	0.08	0.02	0.90
Lithuania	0.12	0.05	0.83
Luxembourg	0.34	0.49	0.17
Netherlands	0.10	0.87	0.03
Poland	0.90	0.03	0.07
Portugal	0.35	0.52	0.13
Spain	0.09	0.87	0.04
Sweden	0.10	0.86	0.04
UK	0.01	0.99	0.00

Table 1. The membership values of the EU countries to the clusters.

Fuzzy clustering generalizes partition clustering methods by allowing an country to be partially classified into more than one cluster. Table 1 gives the fuzzy c-means membership levels for each cluster (all the membership values for each cluster are complementary, in the sense that their sum is one). The membership-values are shown in Table 1, where the highest value appears in bold. Table 2 shows the values of several indices as fuzzy cluster validity measures. The values of these indices can be independently used in order to evaluate and compare clustering partitions or even to determine the number of clusters existing in a data set. Therefore, three is indicated as the optimal number of clusters for all the indices as it corresponds to maximum partition coefficient (pc) and minimum entropy (pe), xie.beni (xb) and fukuyama.sugeno (fs).

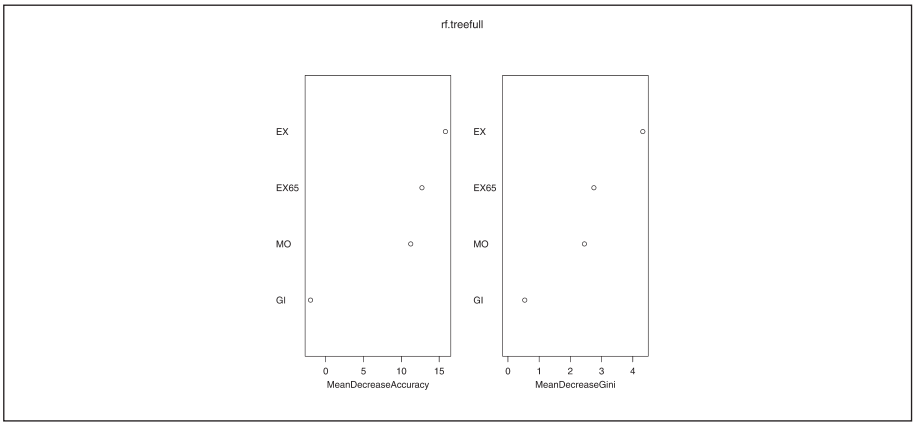


Figure 1. Mortality index importance

index	2	3	4
xb	0.01	0.01	0.03
fs	-26720.47	-26720.92	-18588.22
pc	0.73	0.73	0.61
pe	0.50	0.50	0.73

Table 2. Results of the validation indices by using data of the EU countries. number of clusters

Next step is the characterization of a country group based on the most important mortality indicators using random forests a combination of CART Liaw and Wiener (2002).

2.3. Random forest for mortality indicators in 2009

In order to check the importance of life expectancy at birth, the randomForest package is used. This produces a measure of the importance of the mortality indicator in 2009, and a measure of the proximity of different countries to one another. The Random forest results for accuracy and mortality indicators importance are provided in Figure 1.

3. CONCLUSIONS

In section 2, 21 EU countries were classified by fuzzy c-means cluster analysis, for the time period 1991-2009. As a result for men, one cluster is formed by the western European countries, a second by the Baltic States and a third by Eastern European countries. The clustering method is based on the similarities between mortality surfices which show main mortality time trends. Our results are similar to those obtained by Hatzopoulos and Haberman (2013) who classified 35 countries in west and east clusters, but as we only consider European countries the East-cluster is divided in a different way which in our case gives a Baltic cluster.

Next, random forest was used to implement coherent classification of the countries based on mortality indicators in 2009. For the Baltic States (Estonia, Lithuania and Latvia), the mortality dynamics are distinguishable from the rest of the East-cluster countries, and their life expectancies in 2009 are the lowest ones, less than 70.

ACKNOWLEDGMENTS

This work was funded by a grant from Ministry of Economy and Competitiveness (MTM2013-45381-P).

REFERENCES

Debón, A., Martínez-Ruiz, F., and Montes, F. (2012). "Temporal Evolution of some Mortality Indicators. Application to Spanish Data". *North American Actuarial Journal*, 16 (3): 364-377.

Hatzopoulos, P. and Haberman, S. (2013). "Common Mortality Modeling and Coherent Forecasts. An Empirical Analysis of Worldwide Mortality Data". *Insurance: Mathematics and Economics*, 52 (2): 320-337.

Liaw, A. and Wiener, M. (2002). "Classification and Regression by random-Forest". *R News*, 2 (3): 18-22.

Mackenbach, J. P. (2013). "Convergence and Divergence of life Expectancy in Europe: A Centennial View". *European Journal of Epidemiology*, 28 (3): 229-240.

Spinakis, A., Anastasiou, G., Panousis, V., Spiliopoulos, K., Palaiologou, S., and Yfantopoulos, J. (2011). *Expert Review and Proposals for Measurement of Health Inequalities in the European Union*. Technical report, European Commission Directorate General for Health and Consumers, Luxembourg. ISBN 978-92-79-18529-8.

25. OPTIMAL DESIGNS SUBJECT TO COST CONSTRAINTS IN SIMULTANEOUS EQUATIONS MODELS

V. Casero-Alonso and J. López-Fidalgo
University of Castilla-La Mancha, Department of Mathematics, Institute of Mathematics Applied to Science and Engineering, Spain,
Victormanuel.Casero@uclm.es, Jesus.LopezFidalgo@uclm.es

ABSTRACT

A procedure based on a multiplicative algorithm for computing optimal experimental designs subject to cost constraints in simultaneous equations models is presented. A convex criterion function based on a usual criterion function and an appropriate cost function is considered. A specific L-optimal design problem and a numerical example are taken from Conlisk (1979) to compare the procedure. The problem would need integer non linear programming to obtain exact designs. To avoid this he solves a continuous non linear programming problem and then he rounds-off the number of replicates of each experiment. The procedure provided in this paper reduces dramatically the computational efforts in computing optimal approximate designs. It is based on a specific formulation of the asymptotic covariance matrix of the full-information maximum likelihood estimators, which simplifies the calculations. The design obtained for estimating the structural parameters of the numerical example by this procedure are not only easier to compute, but also more efficient than the design provided by Conlisk.

1. INTRODUCTION

Optimal experimental designs when constraints are imposed have been considered in the literature widely (see e.g. Cook and Fedorov, 1995). In this paper the problem is focused on the total cost of the experiment. A scalar coefficient will tune the

compromise between information and cost (benefit). The novelty is to consider this approach for a model of simultaneous equations. These models are very common in the economics literature among other areas.

As far as the authors know the first to consider the problem of designing experiments for simultaneous equations models was Conlisk (1979). For him the natural approach, as in standard regression, is to minimize a scalar function of the covariance matrix of the estimated structural coefficients of the simultaneous equations model. He proposed a design objective function based either on the three-stage least squares (3SLS) estimators or on the full–information maximum likelihood (FIML) estimators. Moreover the non- linear restrictions on the reduced-form coefficients implied by the structural model causes the design criteria to be functions of unknown parameters. There is not much literature about optimal design for simultaneous equations models. Casero-Alonso and López-Fidalgo (2014) proposed an alternative approach using triangular simultaneous equations models to the optimal experimental design problem for a linear regression model with explanatory variables which are not subject to the control of the practitioner. Here it is extended to a problem with cost constraints. The direct expression of the asymptotic covariance matrix of the FIML estimators provided by Rothenberg and Leenders (1964) is used for a more general model and an efficient iterative procedure is proposed to find optimal designs.

In order to compute optimal designs we consider a multiplicative algorithm based on a convex criterion function. The multiplicative algorithm (e.g. Torsney and Martín-Martín, 2009) is used to update simultaneously all weights of the design points, step by step. A modification of the weighted average of an optimality criteria and a cost function (Pronzato, 2010) is considered in order to better fit the procedure. Cook and Wong (1994) showed that constrained and compound optimal design approaches to handling multiple objectives are equivalent. This approach is based on approximate designs. But an exact design is the final objective.

2. OPTIMAL EXPERIMENTAL DESIGNS IN SIMULTANEOUS EQUATIONS MODELS

A general problem of designing an optimum experiment for a simultaneous equations model subject to cost constraints is formulated.

First the usual notation of simultaneous equations is introduced. Let us suppose a complete system of r linear stochastic structural equations in r jointly dependent variables and k predetermined explanatory variables,

$$y_i = Y_i\alpha_i + X_i\beta_i + u_i = Z_i\delta_i + u_i, \qquad i = 1, ..., r, \tag{1}$$

where yi is the $n \times 1$ vector of observations of one of the endogenous dependent variables, n is the number of experiments, Y_i is the $n \times r_i$ matrix of values formed by the $r_i \leq \mathrm{r} - 1$ explanatory endogenous variables included in that equation, X_i is the $n \times k_i$ matrix of values formed by the $k_i \leq k$ explanatory exogenous variables included in that equation, α_i and β_i are $r_i \times 1$ and $k_i \times 1$ structural parameter vectors to be estimated, for the included endogenous and exogenous variables respectively, and u_i is the $n \times 1$ vector of structural disturbances which have zero mean and are independent and homoscedastic across observations, that is with common covariance matrix, cov$(u_{1h}, ..., u_{rh}) = \Sigma$, $\mathrm{h} = 1, \ldots, n$. Let $y_i = Z_i\delta_i + u_i$ where $Z_i = (Y_i, X_i)$ and $\delta_i = (\alpha_i, \beta_i)$.

Furthermore, let us suppose that the reduced–form of the structural equations system exists, $Y = X\Pi + V$, where Y is the matrix of endogenous variable observations, X is the matrix of values of the exogenous variables, Π is the coefficients matrix, and V is the reduced–form disturbances matrix.

Let $Y_i = X\,\Pi_i$ + Vi where Π_i and V_i are submatrices of Π and V for that part of the reduced–form corresponding to Y_i.

The asymptotic covariance matrix of 3SLS or FIML estimators of the structural coefficients δ of (1) assuming the normal distribution for the disturbances is

$$\text{plim (cov } \hat{\delta}) = \left[-n \lim_{n\to\infty} E\left(\frac{\partial^2 L^{\star}}{\partial\delta\partial\delta^T}\right)\right]^{-1} = \left\{H^T\left[\Sigma^{-1} \otimes (X^T X)\right] H\right\}^{-1}, \tag{2}$$

where plim indicates the probability limit, L^* is the 'concentrated' likelihood function (see Rothenberg and Leenders 1964), $\otimes$ represents the Kronecker product of matrices and

$$H = \begin{pmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_r \end{pmatrix}, \tag{3}$$

where $H_i = (\Pi_i, J_i)$ and J_i is the matrix of unit vectors such that $X_i = X J_i$ selects from X the exogenous variables of the ith equation.

To illustrate the procedure introduced in this paper, a specific L–optimal design problem is considered. An L-optimal design ξ^* is a design which minimizes the trace of $LM(\xi)^{-1}$ (Atwood, 1976), where $M(\xi)$ is the Information Matrix associated with an approximate design ξ and L is a positive definite user–selected matrix. L is chosen to reflect the researcher's interest in the study. Let us suppose that the experimenter's objective function is the sum of estimation error variances for predicting y_i, $i = 1, ..., r$. The cost constraint $\sum_{i=1}^{m} c_j n_j \leq C$ where c_j is the cost of taking one observation at x_j, n_j is the number of times x_j appears in the final design and C is the total budgetof the experiment. Then the specific L–optimal experimental design problem can be stated as

$$\min_{n_1,\ldots,n_m} \text{tr} \left\{ \left[H^T (I_r \otimes \sum_j x_j^T x_j) H \right] \left[H^T \left(\Sigma^{-1} \otimes \sum_j n_j x_j^T x_j \right) H \right]^{-1} \right\} \qquad (4)$$
$$\text{s.t. } \sum_j n_j c_j \leq C \text{ and } n_1 \geq 0, \ldots, n_m \geq 0,$$

where I_r is the identity matrix of order r. As Conlisk (1979) pointed out, though the n_j must be integers (exact design), the usual procedure is to approximate by solving the continuous nonlinear programming problem and rounding–off the optimal values of n_j.

3. AN EFFICIENT ITERATIVE PROCEDURE

A procedure to solve the experimental design problem in a simultaneous equations model subject to a cost constraint stated in (4) is established. It is an alternative to the computationally expensive continuous nonlinear programming problem.

Given that $X^T X$ of (2) can be expressed as $\sum_j n_j x_j^T x_j$ in (4), approximate designs may be used here. Then a multiplicative algorithm for approximate designs can be applied. Therefore, one of the advantages of the procedure described below is the reduction in time of computation (see Section 5). Problem (4) is equivalent to obtaining

the optimal approximate design ξ^* which minimizes the convex criterion function (Cook and Wong, 1994)

$$\Phi_\beta(\xi) = \Phi(\xi) - \beta B(\xi), \tag{5}$$

where ξ is an approximate design, Φ is an optimality convex criterion function (usually a function of $M^{-1}(\xi)$), $B(x_j)$ is an appropriate benefit function based on the known cost function c_j, $B(\xi) = \Sigma_j \xi(x_j)B(x_j)$ which is linear in ξ and $\beta > 0$ is a parameter to be tuned. Note that the minus sign assigned to β is to obtain the maximum benefit/minimum cost.

The General Equivalence Theorem, GET (Kiefer and Wolfowitz, 1959) says that ξ^* is $\Phi\beta$ –optimal if and only if

$$\partial\Phi_\beta(\xi^\star, 1_{x_j}) = \text{tr}\left\{\nabla\Phi(\xi^\star)\left[M(1_{x_j}) - M(\xi^\star)\right]\right\} - \beta\left[B(x_j) - B(\xi^\star)\right] \geq 0, \quad j = 1, ..., m,$$

where $\partial\Phi(\xi, \xi 0)$ is the directional derivative of Φ at ξ in the direction of ξ',

$$\partial\Phi(\xi, \xi') = \lim_{\varepsilon \to 0^+} \frac{\Phi\left[(1-\varepsilon)\xi + \varepsilon\xi'\right] - \Phi(\xi)}{\varepsilon}$$

and 1_{xj} is the design concentrated at x_j, i.e. with weight 1 at x_j. Based on the GET a multiplicative algorithm is used (Torsney and Martín-Martín, 2009). For a specific value of β, at step s + 1 the weights of the design are updated,

$$\xi^{(s+1)}(x_j) = \xi^{(s)}(x_j)\frac{\beta B(x_j) - \text{tr}\left[\nabla\Phi(\xi^{(s)})M(1_{x_j})\right]}{\beta B(\xi^{(s)}) - \text{tr}\left[\nabla\Phi(\xi^{(s)})M(\xi^{(s)})\right]}. \tag{6}$$

For setting a stopping rule, a lower bound of the efficiency of the new design $\xi^{(s+1)}$ is computed. Again it is based on the GET, and it depends on $\Phi_\beta(\xi^{(s+1)})$,

- If $\Phi_\beta(\xi^{(s+1)}) > 0$ then $\text{eff}_{\Phi_\beta}\left[M(\xi^{(s+1)})\right] \geq 1 + K$,
- If $\Phi_\beta(\xi^{(s+1)}) < 0$ then $\text{eff}_{\Phi_\beta}\left[M(\xi^{(s+1)})\right] \geq 1/(1 + K)$,

where $K = \min_{x_j} \partial\Phi_\beta(\xi^{(s+1)}, 1_{x_j})/\Phi_\beta(\xi^{(s+1)})$.

For computing the optimal design ξ^* for the criterion (5) the following procedure is established:

1. For each particular value of β an optimal approximate design is obtained numerically, $\xi^{(\beta)} = \arg\min_{\xi} \Phi_\beta(\xi) = \arg\min_{\xi} [\Phi(\xi) - \beta B(\xi)]$, using the multiplicative algorithm (6).

2. Let $\beta^\star = \arg\min_{\beta} \Phi\left(n^{(\beta)}\xi^{(\beta)}\right)$, where $n^{(\beta)} = C/C(\xi^{(\beta)})$ is the total β number of experiments for each β, and $C(\xi^{(\beta)}) = \sum_j \xi^{(\beta)}(x_j)c_j$.

3. Finally $n^{(\beta^\star)}\xi^{(\beta^\star)}(x_j)$ has to be rounded–off to integers, n^*, in such a way $\sum_j n_j^\star \approx n^{(\beta^\star)}$ and $\sum n_j^\star c_j \approx C$.

This procedure could be adapted to any design problem.

4. AN ILLUSTRATIVE EXAMPLE

To compare the procedure described above with the continuous problem stated in (4) the numerical example of Conlisk (1979) is considered.

A major oil company decides to experiment with controlled variation in the prices of gas and repairs, P_g and P_r, using its service station outlets as a cross–section of observations. Three alternate values (0.93, 1, 1.07) for each price variable (in an appropriate normalized form) are considered. Other controlled variables are the dummies D_g and D_r, describing whether or not trading stamps are offered with gas and repair sales respectively. The design space X is 4–dimensional, (P_g, P_r, D_g, D_r) with values 0 or 1 for D_g and D_r. This yields $3^2 \times 2^2 = 3^6$ experimental possible conditions.

The goal is to estimate the coefficients of the demand model (structural equations) of the quantities of gas and repairs sold, Q_g and Q_r. To solve the corresponding experimental design problem (4) we must supply nominal values for Σ, disturbance

covariance matrix, and H , where H depends on Π via (3) and Π in turn depends on the structural coefficients α_i and β_i:

$$\begin{cases} Q_g = 1.38 + 0.1Q_r - 0.5P_g + 0.04D_g + u_1, \\ Q_r = 0.39 + 0.8Q_g - 0.2P_r + 0.02D_r + u_2, \end{cases} \qquad \Sigma = \begin{pmatrix} 0.1 & 0.06 \\ 0.06 & 0.1 \end{pmatrix}.$$

The cost (in terms of percentages of profit lost by stations during the experiment) of making one experiment at the jth experimental condition (service station) is defined by

$$c_j = 5 + 3673(P_g^{(j)} - 1)^2 + 2857(P_r^{(j)} - 1)^2 + 1633(P_g^{(j)} - 1)(P_r^{(j)} - 1) + 5(1 - D_g^{(j)}).$$

The optimal structural design, provided by Conlisk (1979), is presented in Table 1. This exact design requires 529 experiments with a total cost of 6471.

Table 1 also presents the optimal design obtained with the procedure described in this work. The benefit function considered is $B(\xi) = \Sigma_j \xi(x_j) B(x_j)$ *where* $B(x_j) = 1/c_j$. The values of β are taken from the interval [0, 2000]. Then β* = 586.5. This design requires 446 experiments, 16% less than Conlisk exact design, but the cost is about the same.

5. CONCLUSION

There are two main advantages of the procedure proposed in this paper. On the one hand, it saves time of computations. The time for obtaining a solution of a reduced version of the example is around 60 seconds, whereas our procedure needs only 6 seconds (both implemented in *Mathematica*). On the other hand, the designs obtained with our procedure have much larger efficiencies. To compare the designs we compute the relative L-efficiency: $\text{eff}_L(\xi, \xi') = \frac{\Phi_L[M(\xi')]}{\Phi_L[M(\xi)]}$, where exact designs with integers $n_j, j = 1, \ldots, m$ are considered. The exact design obtained with the procedure is 26.6% more efficient than the structural design provided by Conlisk.

Moreover, the procedure could be adapted to any design problem.

		P_r	0.93		1		1.07	
		D_r	0	1	0	1	0	1
P_g	D_g							
0.93	0		3	1	17	7	6	11
	1		1	0	2	0	1	4
1	0		17	2	49	57	1	6
	1		8	0	137	136	1	18
1.07	0		3	1	0	1	0	0
	1		11	5	5	15	0	3

Conlisk (1979)

		P_r	0.93		1		1.07	
		D_r	0	1	0	1	0	1
P_g	D_g							
0.93	0		4	16	0	0	17	6
	1		0	0	0	0	0	0
1	0		0	0	0	0	0	0
	1		0	0	165	166	0	0
1.07	0		10	27	0	0	26	9
	1		0	0	0	0	0	0

Procedure of this work

Table 1. Optimal designs for estimating the structural parameters.

Another issue is the consideration of approximate designs when the problem considered requires an exact design. We have applied the rounding-off methods included in the software BAZI to the approximate solutions obtained. Therefore, the last step will be to choose the best one with respect to the criteria considered.

ACKNOWLEDGMENTS

This work has been supported by Ministerio de Educación y Ciencia and Fondos FEDER MTM2010–20774–C03–01, MTM2013–47879–C2–1–P and Junta de Comunidades de Castilla-La Mancha PEII10–0291–1850.

REFERENCES

Atwood, C.L. (1976). "Convergent Design Sequences, for Sufficiently Regular Optimality Criteria". *Ann Stat.* 4 (6): 1124-1138.

Casero-Alonso, V., and López-Fidalgo, J. (2014). "Experimental Designs in Triangular Simultaneous Equations Models". *Stat Pap* (in press) doi: 10.1007/s00362-014-0581-y.

Conlisk, J. (1979). "Design for Simultaneous Equations". *J Econometrics*, 11 (1): 63-76

Cook, D., and Fedorov, V. (1995). "Constrained Optimization of Experimental Design". *Statistics*, 26 (2): 129-178.

Cook, R.D., and Wong, W.K. (1994). "On the Equivalence of Constrained and Compound Optimal Designs". *J Am Stat Assoc*. 89 (426): 687-692.

Kiefer, J., and Wolfowitz, J. (1959). "Optimum Design in Regression Problems". *Ann Math Stat*. 30: 271-294.

Rothenberg, T.J., and Leenders, C.T. (1964). "Efficient Estimation of Simultaneous Equations Systems". *Econometrica*. 32: 57-76.

Torsney, B., and Martín-Martín, R. (2009). "Multiplicative Algorithms for Computing Optimum Designs". *J Stat Plan Infer*. 139: 3947-3961.

26. A NUMERICAL ANALYSIS OF RISK MEASURES UNDER PROPORTIONAL REINSURANCE

A. Castañer, M.M. Claramunt and M. Mármol
Universitat de Barcelona, Research Group on Actuarial and Financial Modelling (2014SGR152), Spain,
{acastaner, mmclaramunt, mmarmol}@ub.edu

ABSTRACT

Reinsurance is one of the tools that an insurer can use to mitigate the underwriting risk and then to control solvency. We focus on the proportional reinsurance arrangements and we numerically examine decision problems of the insurer with respect to the reinsurance strategy. To this end, we use as decision tools not only the probability of ruin but also some measures of the random variable deficit at ruin if ruin occurs. The discounted penalty function is employed to calculate all these risk measures as particular cases. We consider the classical risk theory model assuming a Poisson process and that the individual claim amount follows a linear combination of an exponential (β) and an Erlang (2, β), modified with a proportional reinsurance with a retention level that is not constant and depends on the level of the surplus.

1. INTRODUCTION

One of the tools that an insurer has to control solvency is reinsurance. There is a vast bibliography about this subject including also several related optimization problems: Castañer *et al.* (2010, 2012), Dickson and Waters (1996), Centeno (2002, 2005, 1986), Centeno and Simões (2009), Dimitrova and Kaishev (2010), Kaishev and Dimitrova (2006) and Castañer *et al.* (2013), to mention just a few.

In the analysis of solvency, the probability of ruin is the main measure used, but other measures related to the deficit at ruin if ruin occurs are also of interest. We concentrate in the expectation, the Value at Risk *(V aR)* and the Tail Value at Risk *(T V aR)* of this deficit at ruin if ruin occurs. We perform a numerical analysis of all these risk measures in the classical risk theory model considering proportional reinsurance arrangements. In the classical proportional reinsurance the retention level is constant. In Castañer *et al.* (2010, 2012), and Castañer *et al.* (2014), a threshold proportional reinsurance, where the retention level is not constant and depends on the level of the surplus, is defined. This threshold proportional reinsurance includes the classical proportional reinsurance as a particular case.

In the classical risk theory model, the surplus, *R(t)*, at a given time $t \in [0, \infty)$ is defined as $R(t) = u + ct - S(t)$, with $u = R(0) \geq 0$ being the insurer's initial surplus, $S(t)$ the aggregate claims and c the instantaneous premium rate. $S(t)$ is modeled as a compound Poisson process

$$S(t) = \sum_{i=1}^{N(t)} X_i.$$

The claim number process $\{N(t)\}_{t=0}^{\infty}$ is assumed to be Poisson with parameter λ. The random variables $\{X_i\}_{i=1}^{\infty}$ are the positive claim severities, which are i.i.d. random variables with common probability density function $f(x)$ and $\{N(t)\}_{t=0}^{\infty}$ is independent of $\{X_i,\ i \geq 1\}$. We assume that the insurer's premium income is received continuously at rate c per unit time, where $c = \lambda E[X](1+\rho)$, and $\rho > 0$ is the premium loading factor.

The time of ruin is $T = min\ \{t \geq 0 \mid R(t) < 0\}$, with $T = \infty$ if $R(t) \geq 0$ for all $t \geq 0$. The deficit at ruin if ruin occurs is $Y = |R^{+}(T)|$ and the surplus immediately prior to ruin is $R^{-}(T)$. The probability of ultimate ruin from initial surplus u is denoted $\psi(u)$ and defined by

$$\psi(u) = P[T < \infty \mid R(0) = u] = E\{I(T < \infty) \mid R(0) = u\},$$

where $I(\mathrm{A}) = 1$ if A occurs and $I(\mathrm{A}) = 0$ otherwise.

Gerber and Shiu (1998) introduced the Gerber-Shiu discounted penalty function $\phi(u)$,

$$\psi(u) = P[T < \infty \mid R(0) = u] = E\{I(T < \infty) \mid R(0) = u\}, \quad (1)$$

being $\delta \geq 0$ the discounted factor, and $w(l, j)$, $l \geq 0, j > 0$, the penalty function, so that $\phi(u)$ is the expected discounted penalty payable at ruin. This function can be used to study the traditional quantities of interest in classical ruin theory, such as ruin probability, time of ruin or deficit at ruin. Therefore, depending on the penalty function $w(l, j)$, we can obtain different interpretations for the Gerber-Shiu function. For $w(l, j) = 1$, the Gerber-Shiu function equals to the defective Laplace transform of the time of ruin being δ the parameter, and if in addition, we consider $\delta = 0$, the ultimate ruin probability is obtained. For $w(l, j) = j^m$ and $m \geq 1$, the ordinary discounted moments of the deficit at ruin if ruin occurs can be calculated. We also obtain the distribution function of the discounted deficit at ruin if ruin occurs, for $w(l, j) = I(j \leq y)$.

The threshold proportional reinsurance strategy is a dynamic strategy with a retention level that is not constant and depends on the level of the surplus, $R(t)$. A retention level k_1 is applied whenever the reserves are less than a threshold $b \geq 0$, and a retention level k_2 is applied in the other case. Then, the premium income retained is c_1 and c_2, respectively. We consider that the retention levels give new positive security loadings for the insurer, i.e. the net profit condition is always fulfilled. Then, we can define $\rho_1 = \rho_R - \frac{\rho_R - \rho}{k_1}$ and $\rho_2 = \rho_R - \frac{\rho_R - \rho}{k_2}$, being ρ_R the loading factor of reinsurer.

2. THE MODEL

In the classical risk theory model assuming a compound Poisson process for the aggregate claims and a phase-type distribution for the individual claim amount, with threshold proportional reinsurance, the discounted penalty function (1) behaves differently, depending on whether initial surplus u is below or above the level b. Hence, for notational convenience, we write

$$\phi(u) = \begin{cases} \phi_1(u), & 0 \leq u < b, \\ \phi_2(u), & u \geq b. \end{cases}$$

In Castañer *et al.* (2010) the following theorem for the integro-differential equation for the Gerber-Shiu function (1) is obtained.

Theorem 2.1. The discounted penalty function $\varphi(u)$ in a Poisson process model satisfies the integro-differential equations

$$\phi(u) = \begin{cases} \phi_1(u), & 0 \le u < b, \\ \phi_2(u), & u \ge b. \end{cases} \tag{2}$$

where

$$\begin{aligned} \phi_1'(u) &= \frac{\lambda+\delta}{c_1}\phi_1(u) - \frac{\lambda}{c_1}\int_0^{\frac{u}{k_1}} \phi_1(u-k_1x)dF(x) - \frac{\lambda}{c_1}\xi_1(u), \\ \phi_2'(u) &= \frac{\lambda+\delta}{c_2}\phi_2(u) - \frac{\lambda}{c_2}\left[\int_0^{\frac{u-b}{k_2}} \phi_2(u-k_2x)dF(x)\right. \\ &+ \left.\int_{\frac{u-b}{k_2}}^{\frac{u}{k_2}} \phi_1(u-k_2x)dF(x)\right] - \frac{\lambda}{c_2}\xi_2(u), \end{aligned} \tag{3}$$

and

$$\xi_1(u) = \int_{\frac{u}{k_1}}^{\infty} w(u, k_1x-u)f(x)dx, \quad \xi_2(u) = \int_{\frac{u}{k_2}}^{\infty} w(u, k_2x-u)f(x)dx. \tag{4}$$

Let $w(R-(T), |R+(T)|)$ be a non-negative function of $R-(T) > 0$, the surplus immediately before ruin, and $R+(T) > 0$ the surplus at ruin.

We assume that the individual claim amount follows a linear combination of an exponential(β) and an Erlang(2, β). That is a P H (α, S) with $\alpha = (\alpha_1, \alpha_2)$ and

$$S = \begin{pmatrix} -\beta & \beta \\ 0 & -\beta \end{pmatrix}.$$

From Castañer *et al.* (2014), the ordinary differential equations from the integro-differential equation included in Theorem 2.1 are

$$\begin{aligned} \phi_s'''(u) &= \left(\frac{\lambda+\delta}{c_s} - \frac{2\beta}{k_s}\right)\phi_s''(u) \\ &+ \left(\frac{(\lambda+\delta)\,2\beta}{c_sk_s} - \frac{\beta^2}{k_s^2} - \frac{\lambda\alpha_2\beta}{c_sk_s}f(0)\right)\phi_s'(u) + \frac{\delta\beta^2}{c_sk_s^2}\phi_s(u), \end{aligned} \tag{5}$$

where $\phi_s(u)$, $s = 1, 2$ being $s = 1$ for $0 < u < b$ and s = 2 for $u > b$. The function $\xi_s(u)$ is

$$\xi_s(u) = e^{\frac{-\beta u}{k_s}} \left(C_1^{(s)} + C_2^{(s)} + C_3^{(s)} \right), \ s = 1, 2.$$

For the penalty functions

i) $w(l, j) = 1$: $C_1^{(s)} = \alpha_1$, $C_2^{(s)} = \alpha_2$ and $C_3^{(s)} = \alpha_1 \frac{\beta}{k_s}$.

ii) $w(l, j) = j^m$: $C_1^{(s)} = \alpha_1 m! 2 \frac{k_s^m}{\beta^m}$, $C_2^{(s)} = \alpha_2 \frac{m! k_s^m}{\beta^m}$ and $C_3^{(s)} = \alpha_1 m! \frac{k_s^{m-1}}{\beta^{m-1}}$.

iii) $w(l, j) = I(j \leq y)$: $C_1^{(s)} = \alpha_1 \left(1 - e^{-\frac{\beta}{k_s} y}\right)$, $C_2^{(s)} = \alpha_2 \left(1 - e^{-\frac{\beta}{k_s} y}\right) - \alpha_1 \frac{\beta}{k_s} y e^{-\frac{\beta}{k_s} y}$ and $C_3^{(s)} = \alpha_1 \frac{\beta}{k_s} \left(1 - e^{-\frac{\beta}{k_s} y}\right)$.

Then, the Gerber-Shiu function is

$$\phi(u) = \begin{cases} \phi_1(u) = \sum_{i=1}^{3} F_i e^{r_i u}, & 0 \leq u < b, \\ \phi_2(u) = \sum_{i=4}^{6} F_i e^{r_i u}, & u \geq b. \end{cases} \tag{6}$$

with ri, $i = 1, \ldots, 6$, real and distinct roots of the characteristic equations of (5). The values of F_i, $i = 1, \ldots, 6$, are determined from six conditions (see Castañer *et al.* (2014), for more details). To obtain the ruin probability, and the moments and the deficit at ruin (not their present values) we have to consider $\delta = 0$, so $r_3 = r_6 = 0$.

3. SOME RESULTS

In the numerical analysis, we assume that $\beta = 2$, $\alpha = (0.5, 0.5)$, $\lambda = 1$, $\rho = 0.4$, $\rho_R = 0.5$ and $\delta = 0$. We are interested in finding, if it exists, the best strategy for the insurer in order to minimize its ruin probability. We solve two optimization problems. In the first one, the insurer only considers the proportional reinsurance option, while in the second one, the insurer also considers the threshold proportional reinsurance option.

In Tables 1 and 2, we include the result of this minimizations and the comparison with the probabilities of ruin without reinsurance, for different values of the initial capital u. In addition, for the optimal threshold reinsurance strategies, in Table 3, the

expectation, the variance and the Value at Risk and the Tail Value at Risk at level 99.5% of the deficit at ruin if ruin occurs are shown.

u	(b, k_1, k_2) Threshold	k Proportional
0	(0.403104, 1, 0.351626)	1.00000
0.25	(0.403102, 1, 0.351625)	0.47684
0.5	(0.403968, 1, 0.354007)	0.41151
1	(0.405915, 1, 0.359265)	0.38698
2	(0.407031, 1, 0.362216)	0.37590
3	(0.407427, 1, 0.363264)	0.37235
5	(0.407757, 1, 0.364128)	0.36956

Table 1. Optimal points.

u	$\psi_{TH}(u)$	$\psi_P(u)$	$\psi(u) - \psi_{TH}(u)$	$\psi(u) - \psi_P(u)$
0	0.629977	0.714286	0.084309	0.000000
0.25	0.460979	0.550169	0.122815	0.033625
0.5	0.317820	0.379671	0.153875	0.092024
1	0.147567	0.177239	0.157640	0.127968
2	0.031802	0.038294	0.095360	0.088868
3	0.006852	0.008258	0.046111	0.044706
5	0.000318	0.000384	0.008870	0.008804

Table 2. Minimum probabilities of ruin with reinsurance.

u	$E[Y]$	$V[Y]$	$VaR99.5\%$	$TVaR99.5\%$
0	0.326255	0.097976	1.613070	1.890450
0.25	0.298149	0.0905538	1.562500	1.839210
0.5	0.284289	0.0875708	1.543070	1.820110
1	0.283708	0.0872991	1.541650	1.818690
3	0.283076	0.0870389	1.540250	1.817300

Table 3. Some measures of the deficit at ruin if ruin occurs.

REFERENCES

Castañer, A., Claramunt, M.M., and Lefèvre, C. (2013). "Survival Probabilities in Bivariate Risk Models, with Application to Reinsurance". *Insurance: Mathematics and Economics*, 53, 632-642.

Castañer, A., Claramunt, M.M., and Mármol, M. (2010). "Deficit at Ruin with Threshold Proportional Reinsurance". *Insurance Markets and Companies: Analyses and Actuarial Computations*, 1, 38-44.

Castañer, A., Claramunt, M.M., and Mármol, M. (2012). "Ruin Probability and Time of Ruin with a Proportional Reinsurance Threshold Strategy". *TOP*, 20, 614-638.

Castañer, A., Claramunt, M.M., and Mármol, M. (2014). "Some Optimization and Decision Problems in Proportional Reinsurance". *UB Economics Working Papers. Col·lecció d'Economia,* E14 (310), 1-28.

Centeno, L. (1986). "Measuring the Effects of Reinsurance by the Adjustment Coefficient". *Insurance: Mathematics and Economics*, 5, 169-182.

Centeno, L. (2002). "Measuring the Effects of Reinsurance by the Adjustment Coefficient in the Sparre Anderson Model". *Insurance: Mathematics and Economics*, 30, 37-49.

Centeno, L. (2005). "Dependent Risks and Excess of Loss Reinsurance". *Insurance: Mathematics and Economics*, 37, 229-238.

Centeno, M.L., and Simões, O. (2009). "Optimal Reinsurance". *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 103, 387-404.

Dickson, D.C.M., and Waters, H.R. (1996). "Reinsurance and Ruin". *Insurance: Mathematics and Economics*, 19, 61-80.

Dimitrova, D.S., and Kaishev, V.K. (2010). "Optimal Joint Survival Reinsurance: An Efficient Frontier Approach". *Insurance: Mathematics and Economics*, 47, 27-35.

Gerber, H.U., and Shiu, E.S.W. (1998). "On the Time Value of Ruin". *North American Actuarial Journal*, 2, 48-72.

Kaishev, V.K., and Dimitrova, D.S. (2006). "Excess of Loss Reinsurance under Joint Survival Optimality". *Insurance: Mathematics and Economics*, 39, 376-389.

27. MODELING LONGEVITY RISK WITH GENERALIZED DYNAMIC FACTOR MODELS AND VINE-COPULAS

H. Chuliá
Universitat de Barcelona, Department of Econometrics and Riskcenter-IREA, Spain,
hchulia@ub.edu

M. Guillén
Universitat de Barcelona, Department of Econometrics and Riskcenter-IREA, Spain,
mgillen@ub.edu

J.M. Uribe
Universidad del Valle, Department of Economics and Universitat de Barcelona, Colombia,
orge.uribe@correounivalle.edu.co

ABSTRACT

We present a methodology to forecast specific mortality rates and estimate longevity and mortality risks. It is based on Generalized Dynamic Factor Models, fitted on the differences of the log-mortality rates. We compare its forecasting performance with models previously proposed in the literature, such as the traditional Static Factor Model fitted on the level of log-mortality rates. We also construct risk measures such as VaR and Tail-VaR by means of vine-copula simulations, taking into account the dependence between the idiosyncratic components of mortality rates. This methodology is applied to the projections of mortality rates of the United Kingdom and the estimation of longevity and mortality risks for hypothetical populations of males and females.

1. INTRODUCTION

Longevity risk, understood as a downward deviation of the population mortality rates from a forecasted mortality trend, has become a topic of great academic interest in recent times, specifically for the actuarial literature (Richards and Jones, 2004). The central point in understanding longevity risk is related to the stochastic nature of mortality rates. As such, their forecast is subject to uncertainty and statistical confidence judgments. This stochastic modeling of mortality rates has been well documented (Cairns *et al.,* 2011) and recommended by the regulators (Continuous Mortality Investigation, 2004). Among the available alternatives to forecast and modeling mortality rates, factor models are an attractive approach, given the low frequency of the mortality data and the relative high number of specific mortality rates to be forecasted. The strategy consists on making the mortality rates dependent on a few numbers of unobserved stochastic factors. By doing this, the number of estimated parameters in the model is significantly reduced and optimal first-efficient forecasts are possible.

In this article we propose a methodology that accommodates advances in two different fronts, estimation and prediction based on Generalized Dynamic Factor Models, and the construction of alternative confidence scenarios by means of vine-copula simulations. In this way, we provide a robust alternative to measure longevity and mortality risks, using risk measures such as Value at Risk, or Tail-Value at Risk. The methodology is applied to forecast mortality rates and to estimate risk measures for the United Kingdom, using data from 1950 to 2011, provided by the Human Mortality Database.

2. METHODOLOGY

Following Bai and Ng (2008), let N be the number of cross-sectional units and T the number of time series observations. In our case, we have 220 cross-sectional units (mortality rates for ages from 0 to 109+ years, for males and females). If we consider males and females separately, then $N = 110$. For $i = 1, \ldots, N$ and $t = 1, \ldots, \mathrm{T}$, the Static Factor Model (SFM) is defined as:

$$x_{it} = \lambda_i F_t + e_{it} \text{ or } \quad x_{it} = C_{it} + e_{it}, \tag{1}$$

where e_{it} is referred to as the idiosyncratic error and λ_i is referred to as the factor loadings. This is a vector of weights that unit i puts on the corresponding r static common factors F_t . $C_{it} = \lambda_i F_t$ refers to the common component of the model. can be a dynamic vector process. i.e. a VAR process. Note that if we set $L = 1$ and $\mathrm{A}(1) = I_r$ and require that $r = 1$, we are in presence of the popular model proposed by Lee and Carter (1992). The static model can be compared with the dynamic model, defined as:

$$x_{it} = \lambda_i(L) f_t + e_{it}, \tag{2}$$

$\lambda_i(L) = (1 - \lambda_{i1}L -, \ldots, -\lambda_{is}L^s)$ where is a vector of dynamic factor loadings of order s. In the case when is finite, we refer to it as a Dynamic Factor Model (DFM), whereas a Generalized Dynamic Factor Model (GDFM) allows s to be infinite. Stock and Watson (2010) provide examples of the former case and Forni *et al*. (2000) introduce the latter. The GDFM differs from traditional static and dynamic models because it places smaller weights on variables having larger idiosyncratic components, so that the idiosyncratic error contained in the linear combination is minimized.

Estimation of the factors use Principal Components, or Singular Value Decomposition, which impose the normalization that $\Lambda'\Lambda/\mathrm{N} = I_r$ and $F'F$ being diagonal (enough to guarantee identification).

Following Stock and Watson (2006), the h-step ahead forecast can be performed directly by projecting x_{it+h} onto the estimated factors lagged h periods, that is, by estimating β in the following equation:

$$X_{t+h} = \beta F_t + e_{t+h}. \tag{3}$$

Unknown factors can be replaced by their (consistent) estimations F_t. In the context of GDFM, these factors are weighted according to the spectral density matrix of the series.

Finally, the dependence arising in the high frequency of the spectra is key to the estimation of unexpected movements in the time series. We approach it by means of copulas, and in this way, we are able of constructing risk measures based on

percentiles of the simulated density[1]. When there are several dimensions involved, as in our case, the direct estimation of a N-dimensional copula is not recommended. Instead, the literature has developed an alternative estimation and simulation procedure, based on bivariate-conditional-copulas as described by Aas *et al.* (2009). This method allows us to estimate percentiles of the multivariate density of the idiosyncratic components in the model, and therefore, to construct risk measures based on specific percentiles of the forecasted error, such as: VaR and $TVaR$. The VaR is the α-quantile of a random variable (in our case a mortality rate), that is $VaR_{\alpha\,(x)} = inf\,\{x \mid F_x(x) \geq \}\alpha$, where F_x is the distribution function of x and α is the significance level $\alpha \in (0,1)$. A complementary measure, the corresponds to the mathematical expectation beyond VaR.

We provide estimations of VaR and $TVaR$ for different configurations of the exposed population. We consider populations of 30,000 males or females ranging from 0 to 109+ years, 18 to 64 years and 65 to 109+ years, as illustrative examples. In the interest of space, we report only the first results.

3. RESULTS AND DISCUSSION

We use of data from 1950 to 2001 to estimate the models and from 2002 to 2011 to measure their relative performance. The Mean Squared Forecasting Error (MSFE) for the changes in log-motility rates for several forecasting horizons is estimated, for males and females, as a weighted average of the individual forecasting errors, the weights being the population in each category for the cases considered in the exercise (see Figure 1).

[1] A comprehensive treatment by McNel *et al.* (2005).

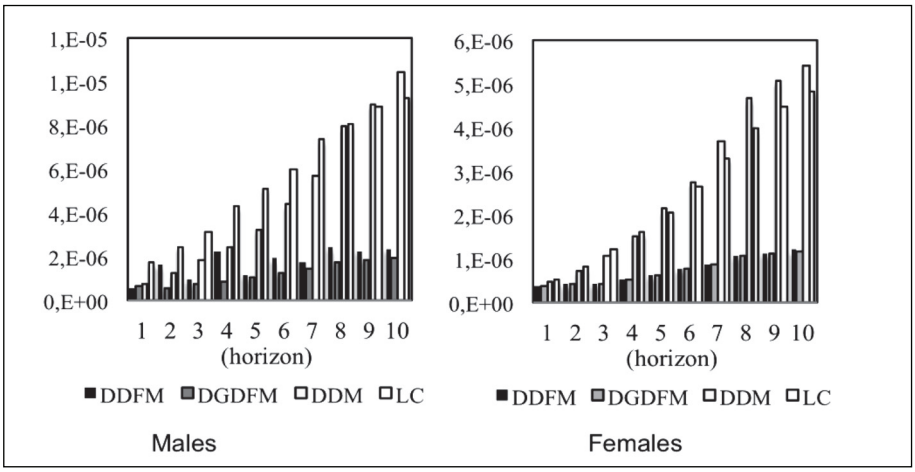


Figure 1. MSFE by Forecasting Horizon. Bars from left to right are in each group are: DFM and GDFM on the log-differences, DFM and SFM-Lee-Carter, on the log-levels.

In general, the GDFM performs better than the other models in forecasting, especially for medium-term horizons. The worst model is the Lee-Carter Model, followed by the DFM in log-rates. Our intuition for this finding is that the log-rates models impose a cointegration relationship between the mortality rates that can be false for the 110 rates as a whole. This in turn could deteriorate the forecasting exercise, especially for medium or long-term horizons. Additionally, we find that the adjustment, through the spectral density matrices applied in the GDFM improves the forecast, as expected, but the improvement is small in magnitude.

Lastly, we report the VaR and $TVaR$ at 0.5% (left-tail) and the same statistics at 99.5 percentiles (right-tail), using several copulas. When we observe the VaR or $TVaR$ at 0.5% we are concerned with longevity risk and, conversely, at the right tail (99.5%) we are referring to the risk of mortality. In both cases risk has to be understood as a departure from the expected number of deaths (ED), forecasted with the best available model (i.e. the GDFM).

Males					
	TVaR	VaR	ED	VaR	TVaR
0-109+	238	239	252	261	262
18-64	78	78	80	82	82
65-109+					
Females					
0-109+	TVaR	VaR	ED	VaR	TVaR
18-64	254	255	259	263	263
65-109+	52	52	54	56	56
	1242	1244	1257	1269	1270

Table 1. Expected number of deaths (ED), VaR and TailVaR Males TVaR VaR ED VaR TVaR.

The forecasting results (ED in Table 1) can be read as the expected number of events (deaths) in one year. In the case of 30,000 males between 0-109+ years, in a population-portfolio that mimics the population structure of the UK males in 2011, we expect to see 252 events (deaths), while we expect to see 259 events in a similar portfolio composed only by woman.

In terms of longevity risk we find that, as expected, longevity risk arises as the age advances. For example, the 99.5% VaR for males between 0 and 109+ years is 261 and the VaR at 0.5% is 239. This indicates that it is possible to assert with a 99.5% of statistical-confidence, in one year, that no more than 261 persons will die, or conversely that no less than 239 will die, even when one expect 252 to die. These calculations increase significantly for the upper ages, for males and females, but at a higher rate for the former.

These results highlight an empirically important finding of our exercise, which is the asymmetric nature of mortality and longevity risks. By using the VaR at 99.5%, longevity risk is 20.8% higher than mortality risk for the portfolio of males (8% in females). Using the $TVaR$s those numbers are 15.8% and 15.3% respectively.

4. CONCLUSION

Introducing dynamics by means of adding lags (new factors) to forecast mortality rates generates a better fit of the models to data, especially when modeling male

populations, although it is the differentiation of the series, which increase considerably the forecasting capability of the factor models. We provide some intuition for these findings and we show that the gains in terms of forecasting are bigger as the forecast horizon increases. The intuition comes from the fact that traditional factor models on the log-mortality rates, as the Lee-Carter's model, impose a cointegration relationship among all the series in the system by assumption, which is very unlikely to be observed in the data, given the high number of series in this kind of exercises. Further improvements can be achieved by using (one sided) generalized principal components in the estimation stage.

It has to be noticed, however, that the models in differences perform worse than the models in levels for short forecasting horizons at ages above 95 years (we do not report these results here). This finding has to do with the fact that the bigger fitting of the models in differences comes at the expense of some degree of over-fitting for the elder ages, especially for males. This over-performance of the models in levels disappears as the forecast horizon increases and, indeed, for 10 years ahead it completely reverses.

Finally, we found that longevity risk is bigger for older populations and particularly higher for men. There is also an asymmetric relation between longevity and mortality risks, which makes it difficult to try to "cancel out" each risk with another in pragmatic terms.

REFERENCES

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). "Pair-Copula Constructions of Multiple Dependence". *Insurance: Mathematics and Economics*, 44 (2), 182-198.

Bai, J., and Ng, S. (2008). "Large Dimensional Factor Analysis". *Foundations and Trends in Econometrics*, 3 (2), 89-163.

Cairns, A., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2011). "Mortality Density Forecasts: An Analysis of six Stochastic Mortality Models". *Insurance: Mathematics and Economics*, 48 (3), 355-367.

Continuous Mortality Investigation (2004). *Projecting Future Mortality : A Discussion Paper*. Working paper no. 3. Institute and Faculty of Actuaries.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). "The Generalized Dynamic-Factor Model: Identification and Estimation". *The Review of Economics and Statistics*, 82 (4), 540-554.

Lee, R.D., and Carter, L.R. (1992). "Modeling and Forecasting U.S. Mortality". *Journal of the American Statistical Association*, 87 (419), 659-671.

Mcneil, A.J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press.

Richards, S., and Jones, G. (2004). *Financial Aspects of Longevity Risk*. Staple Inn Actuarial Society Seasonal Meeting.

Stock, J.H., and Watson, M.W. (2006). "Forecasting with many Predictors". *Handbook of Economic Forecasting* (1), 515-554. Elservier.

Stock, J.H., and Watson, M.W. (2010). "Dynamic Factor Models". *Handbook of Economic Forecasting*. 35-59. Oxford University Press.

28. PERFORMANCE OF THE CLAUSET-SHALIZI-NEWMAN METHOD FOR POWER-LAW FITTING AND COMPARISON WITH OTHER METHODS

Á. Corral
Centre de Recerca Matemàtica, Barcelona, Spain, acorral@crm.cat

I. Serra
Centre de Recerca Matemàtica, Barcelona, Spain, iserra@crm.cat

ABSTRACT

The Clauset-Shalizi-Newman method to find the value of a random variable X above which is well described by a power-law tail is put under scrutiny using simulated data with power-law tails. The effect of considering different shapes in the part of the distribution describing the transition to the power-law tail is explored, as well as the influence of the number of data. Comparison with other methods is done, included a method based on Hill estimator.

1. INTRODUCTION

Power-law distributions, or, more generally, power-law tails (also referred to as Pareto tails) have found an increased range of applicability in the last decades. The reasons for such abundance are diverse: the ubiquity of Zipf's law in linguistics, demography, economy and communications, Newman (2008); the emergence of fractal and critical structures in complex systems, Bak (1996); and last but not least, limiting behavior of threshold exceedances in extreme events (Coles, 2001). In all cases, power-law tails are a sign of the existence of a non-null risk of extremely large events.

2. ANALYSIS

Given a data set, the fitting of a power-law distribution is straightforward, and does not involve more statistical tools than the fitting of any other simple distribution. However, in practice, one is more interested instead in fitting a power-law tail, i.e., one wants to test the hypothesis that there exist a critical value a of the random variable X beyond which

$$P[X > x|X > a] = \left(\frac{a}{x}\right)^{\alpha}$$

to a good approximation, with the exponent $\alpha > 0$ and the truncation parameter $a > 0$, and with a arbitrarily large with respect to the smallest value of X in the data set. The main problem is that different values of a lead to different subsets of the data (as values of X below a are discarded), and therefore the results of the fit for each a cannot be directly compared.

Visual inspection of the plot of the distribution in log-scale (either for the estimated survivor function or the probability density) can be helpful to identify a plausible value of a, but one would be more confident with an automatic method, devoid of any arbitrariness. Other methods, such the use of CV-plot, see Castillo *et al.* (2014), also rely on user visualization.

Recently, Clauset *et al.* (2009) have introduced a protocol to systematically find the proper value of a for power-law tails. In essence, the method explores all possible values of a. For each a, the exponent α is estimated by maximum likelihood, and then a sort of goodness-of-fit (but not goodness-of-fit test) is calculated by the Kolmogorov-Smirnov (KS) statistic. One obtains in this way the KS statistic as a function of a. The key point of the Clauset-Shalizi-Newman method is the assumption that the value of a that signals the starting of the power-law tail is given by the one that minimizes the KS statistic. This can be understood as arising from the fact that, while the data is indeed power-law distributed, the decrease in a will lead to a decrease of the KS distance (because this goes as the inverse of the square root in the number of data, and the number of data increases with a decreasing). On the other hand, as long as the data starts to deviate from the power law, the KS statistic would increase again, due to the bad fit provided by the region outside the power-law range.

Once the candidate value of a is found (with its corresponding estimated value of the exponent α) one needs to test this result. This is done by simulating synthetic samples as close as possible to the original data: boot-strapping the data for X below a and simulating a power law with the obtained exponent for $X > a$. These synthetic samples are submitted to the very same process as the empirical data, with the result, for each synthetic sample, of a KS statistic-function of a; minimization will yield the resulting estimation of the value of a and the exponent in each case. Naturally, these values should be more or less close to the empirical ones, although distinct, in general. For different synthetic samples we will gather a collection of minimum values of the KS statistic. These KS statistic values are the ones that define the distribution with which the empirical minimum KS value is compared, and allow to obtain a p–value, from which the power-law hypothesis is rejected or not.

The Clauset-Shalizi-Newman method has become very popular, in part because it provides a practical and unambiguous recipe to establish or disregard the existence of a power-law tail in empirical data. However, a few researchers have criticized it, based on two main facts: (i) there is no theoretical justification on why the KS statistic should be minimum just at the point in which the data deviate from a power-law behavior (despite the argument above on the dependence on the number of data and the deviations outside the true power-law regime) and (ii) it has been evaluated on data distributed according to a real, simulated power-law tail joined to a non-power-law body and the method has not identified correctly the transition point, yielding a smaller value of a and therefore a p-value equal to zero (for data that, remember, have a real power-law regime). However, this analysis has been performed just for a single case.

An alternative variation has been introduced by Peters *et al.* (2010) and Deluca and Corral (2013), using a more standard procedure. Given an empirical data set, for each possible value of a, a maximum-likelihood fit is performed and the KS statistic is obtained. But the difference starts in the fact that a p–value is obtained for each of these fits. This is done by simulating the power-law tail, for $X > a$, and by applying the same maximum-likelihood estimation and calculation of the KS statistic for each synthetic data set. The comparison of the empirical KS statistic with the KS statistic of the simulated sets yields a p–value for each value of a. The onset of the power-law distribution should be reflected in the behavior of the p–value as a

function of a, transiting from a nearly zero value to a uniform distribution as a crosses from below the transition point. Note that in this case one should pay attention to the multiple-testing nature of this method.

A third method to compare consists in a semi-parametric tool for optimal threshold estimation obtained from non-parametric techniques of bias-corrected for high quantiles estimation. The method is based on Hill estimator under the supposition of heavy tails, see Gomes and Pestana (2007) and provides a consistent estimator of the second order parameters, see Degen and Embrechts (2008).

3. CONCLUSION

In this work we perform an in-depth and exhaustive study of the performance of the Clauset-Shalizi-Newman method, comparing it with the recipe of Peters *et al.* (2010) and Deluca and Corral (2013) and with the more elaborate procedure of Gomes and Pestana (2007). We apply the three methods to different simulated distributions, varying the values of the parameters of the distributions and the size of the data sets. Special care is devoted to the modeling of the crossover region from the body of the distribution to the power-law tail. The results show in which cases the Clauset-Shalizi-Newman method yields reasonable results and when leads to failure, as well as which is the most suitable method, given the empirical and computational restrictions.

ACKNOWLEDGMENTS

Research expenses were founded by projects FIS2012-31324 and MTM2012-31118 from Spanish MINECO, 2014SGR-1307 from AGAUR, and the Collaborative Mathematics Project from La Caixa Foundation.

REFERENCES

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality.* Copernicus, New York.

Castillo, J. del, Daoudi, J., and Lokhart, R. (2014). "Methods to Distinguish Between Polynomial and Exponential Tails". *Scandinavian Journal of Statistics,* 41, 382-393.

Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009). "Power-Law Distributions in Empirical Data". *SIAM Review,* 51, 661-703.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer, London.

Degen, M., and Embrechts, P. (2008). "EVT-based Estimation of Risk Capital and Convergence of High Quantiles". *Advances in Applied Probability,* 40, 696-715.

Deluca, A., and Corral, A. (2013). "Fitting and Goodness-of-Fit Test of Non-Truncated and Truncated Power-Law Distributions". *Acta Geophysisica,* 61, 1351-1394.

Gomes, M.I., and Pestana, D. (2007). "A Sturdy Reduced-Bias Extreme Quantile (VaR) Estimator". *Journal of the American Statistical Association,* 102, 280-292.

Newman, M.E.J. (2005). "Power Laws, Pareto Distributions and Zipf's Law". *Contemporary Physics,* 46, 323-351.

Peters, O., Deluca, A., Corral, A., Neelin, J.D., and Holloway, C.E. (2010). Universality of Rain Event Size Distributions". *Journal of Statistical Mechanics,* P11030.

29. DPLN: DATA ANALYSIS IN INSURANCE AND FINANCE

J. Daoudi
Universitat Pompeu Fabra,Department of economics and business, Spain,
jalila.daoudi@upf.edu

R. Lillo
Universidad Carlos III de Madrid, Department os statistis, Spain,
lillo@est-econ.uc3m.es

ABSTRACT

Operational risk and portfolio risk data exhibit heavy tails properties which in natural way call for an EVT analysis, many authors suggest use the extreme value theory however this approach results in an extremely unrealistic amount of capital at risk. An alternative approach is proposed the DPLN model. This model has been applied to insurance field see Ramirez *et al.* (2008), we compare the DPLN model to others proposed in the literature. An example of operational risk losses is provided.

1. INTRODUCTION

It is often assumed that claim sizes and risk losses are independent and distributed heavy tailed random variables. The extreme value theory does not capture the body of the distribution. The exponential model fit is even worse, it captures neither the body nor the tail. We propose the Double Pareto Lognormal (DPLN) introduced by Reed and Jorgensen (2004).

$$f(x;\alpha,\beta,\tau,\nu) =$$

$$\frac{\alpha\beta}{\alpha+\beta}\frac{1}{x}\phi\left(\frac{logx-\nu}{\tau}\right)\left[R(\alpha\tau-(logx-\nu)/\tau)+R(\beta\tau+(logx-\nu)/\tau)\right],$$

$$R(z) = \Phi^c(z)/\phi(z)$$

is the Mill's ratio, where $\Phi^c(z) = 1 - \Phi(z)$ and $\phi(z)$ and $\Phi(z)$ are the standard normal density and cumulative distributions respectively.

The DPLN distribution does not possess a moment generating function in closed form. However, if $n < \alpha$, the moment of order n exists.

Reed and Jorgensen (2004) show that the DPLN model is a very flexible and consider a various approaches to classical inference. However they not certain problems with the maximum likelihood convergence. We use a bayesian approach to estimate the parameters of DPLN model. On the other hand, we propose The DPLN model to model the aggregate loss distribution.

At the end, we consider 125 operational risk recollected from different banks, and we compare our results to other models proposed in the literature.

2. MODELING THE AGGREGATE DISTRIBUTION

In order to estimate the aggregate loss distribution, we first fit the frequency of occurrence and severity and then performs a process of aggregation or convolution of both. A capital at Risk is then calculated as the quantile 99.9% of the capital estimates for this aggregate distribution in holding period. The most frequent discrete parametric distributions for modeling the frequency of operational risk losses are Poisson distribution and the negative binomial distribution.

The operational risk losses data presents heavy tails. In literature, many distributions have been proposed to model operational risk severity. The route via Subexponential family is a very natural one see Embrechts *et al.* (1997), in particular, the Pareto distribution which has been long used as a model for the tails in several

fields such as hydrology, insurance, finance and environmental science, see for example Degen *et al.* (2007). However the Pareto distribution provides a good model for heavy tails, it is inappropriate for modeling the center of the distribution of many real data sets. Thus, in operational risk results to an inadequate amount of capital with no economic interpretation (more than 100% of the asset size), infinite expected losses and unstable estimates of capital at Risk.

The medium distributions as Gamma distribution and exponential distribution provide stability of capital see Degen *et al.* (2007). Nevertheless, they underestimate the capital at Risk. On the other hand, EVT is very useful when it comes to analysing extreme events but in no way EVT is able to predict exceptional operational risk losses such as those present in the Baring case.

Therefore alternative models are needed to these distributions. Dutta and Perry (2006) proposed the parametric g-h distribution which is well adopted to operational risk more than the extreme value theory, and it results in consistently reasonable capital estimates. Nevertheless, the g-h model is sensitive to the introduction of new data especially if under some scenarios the kurtosis's parameter is a polynomial. On the other hand, in some research the body of the distribution is fitted with the lognormal distribution and the tail by GPD distribution. However, the most difficult problem is the choose of the right threshold which is both, not too low to cause significant bias and not high to lead to significant variance. The bad threshold can lead to under/ overestimate the VaR. In this context, we use the double Pareto lognormal distribution (DPLN). DPLN distribution has a lognormal body and features a power law in both the lower and the upper tail.

3. APPLICATION

The data used for the analysis was collected for several banks participating in the survey to provide individual gross operational losses above a threshold, stating on 2002 to 2007. This data was grouped in eight standardised business lines and seven event types. The data set used here correspond to the 125 largest losses associated with the business line 'Retail Banking' and the event type 'Internal Fraud'.

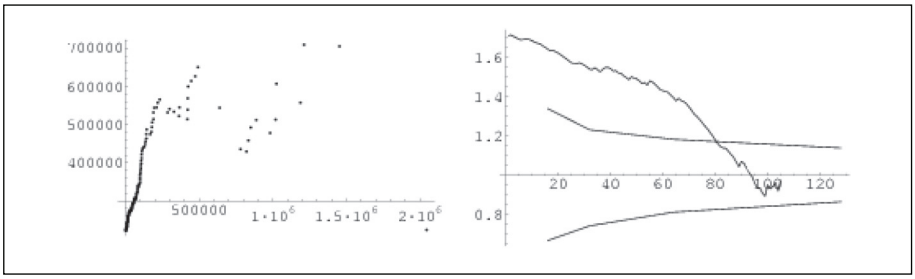


Figure 1. In the left, the ME-plot shows increasing line tendence. In the right, the CV-plot is outside the range defined under exponentiality.

The frequencies of events is modeled by Poisson distribution, then we calculate the capital at Risk under the Pareto distribution and the DPLN distribution.

In case of Pareto distribution, although the fluctuation in the estimated parameters appears relatively small, this leads to a significant capital at Risk estimates as we see in figure 2.

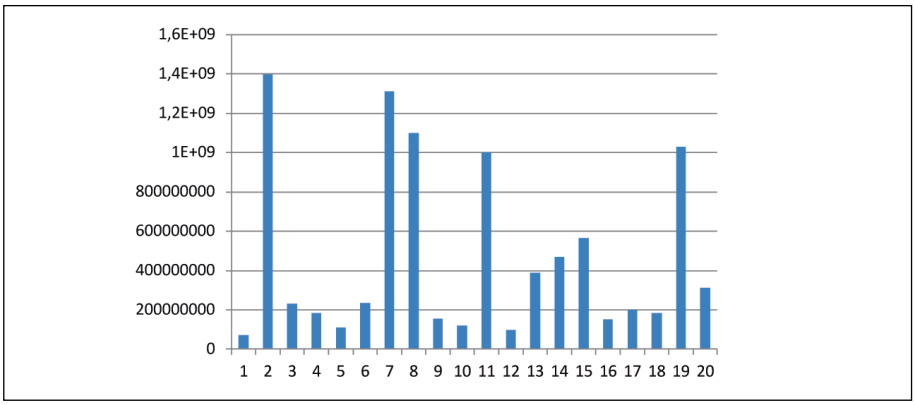


Figure 2. Instability of capital at Risk of Pareto using bootstrapping method.

In Figure 3, we can see that The DPLN distribution gives the most reasonable with economic sense and a stable capital at Risk estimates.

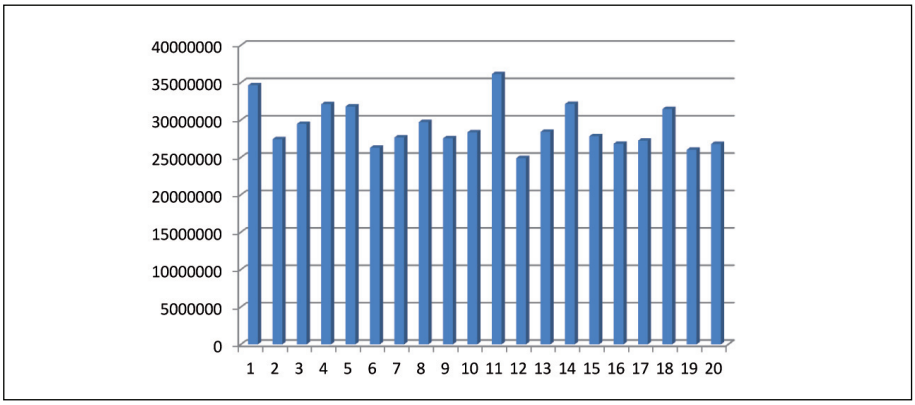


Figure 3. Stability of capital at Risk of DPLN using bootstrapping method.

REFERENCES

Degen, M., Embrechts, P., and Lambrigger, D. (2007). "The Quantitative Modeling of Operational Risk: Between g-and-h and EVT". *ASTIN Bulletin International Actuarial Assotiation-Brussels, Belgium*, 37 (2), 265-291.

Dutta, K., and Perry, J. (2006). "A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital". *FRB of Boston Working Paper*, 6-13.

Ramirez, P., Lillo, R., Wiper, M., and Wilson, S. (2008b). "Inference for Double Pareto Lognormal Queues with Applications". *Statistics and Econometrics Working Papers*, ws084613, Universidad Carlos III de Madrid.

Reed, W., and Jorgensen, M. (2004). "The Double Pareto-Lognormal Distribution: A New Parametric Model for Size Distributions". *Communications in Statistics: Theory and Methods*, 33 (8), 1733-1753.

30. MODELING THE IBEX-35 INDEX USING A COMPOUND POISSON. STOCHASTIC PROCESS

A. Debón
Universitat Politècnica de Valencia, Centro de Gestión de la Calidad y del Cambio, Spain, andeau@eio.upv.es

J.C. Cortés
Universitat Politècnica de Valencia, Instituto Universitario de Matemática Multidisciplinar, Spain, e-mail: jccortes@imm.upv.es

O. Monzó-Chafer
Universitat Politècnica de Valencia, Spain, osmoncha@alumni.upv.es

F. Rodríguez-Santos
Universitat Politècnica de Valencia, Spain, fabianrodriguezsantos@gmail.com

ABSTRACT

The IBEX-35 is the main stock market index in Spain and it is based on the 35 most liquid companies listed on the Spanish stock market. Due to the complexity and inherent uncertainty associated to the social and economic factors that determine the IBEX-35, stochastic models are required to describe its dynamics. In this paper, a mixed approach based on time series techniques and a compound Poisson process is presented to model the IBEX-35. The study has been conducted using weekly closing prices of the stock market from January 1st, 1998 to December 31st, 2012.

1. INTRODUCTION

The functions of a stock exchange are: (1) bring together savers and investors, private and public participants; (2) provide liquidity for investments, and (3) establish security prices on the basis of supply and demand factors. Nowadays, the main function of a stock market is speculative and it is useful for forecasting the health of the economy of its country. The IBEX-35 is the reference stock market index in Spain. The index comprises the 35 most liquid Spanish companies weighted by market capitalization. The goal of modeling the dynamics of the IBEX-35 is to forecast the price behavior in the Spanish stock market, reducing the uncertainty for experts and investors in their decision-making process.

2. DATA ADJUSTMENT

The model uses historical weekly closing prices of the IBEX-35 index from 1998 to 2012. The following adjustments have been applied to the data: (1) adjustment to 52 weeks per annum. Those years with 53 weeks have been adjusted by taking as the last week, the arithmetic average of the last two weeks of the year; (2) applying natural logarithms (Ln) to smooth the data, and model the series with a normal distribution; (3) creating an outlier removal filter according to the Box-Whisker graph criterion $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$. A summary with the descriptive statistics of the raw and adjusted data is shown in Table 1.

	IBEX-35	Ln IBEX-35
Observations	780	780
Mean	9,830	9.167
Median	9,680	9.178
Variance	5,108,883	0.052
Std. Deviation	2,260	0.228
Kurtosis	-0.075	-0.418
Skewness	0.0565	0.026

Table 1. Ln IBEX-35 Descriptive Statistics.

3. METHODOLOGY

The methodology of the model is based on time series analysis. It relies on four components: Trend, Cyclical, Seasonal, and Irregular. Figure 1 shows the breakdown of the adjusted data in these components. The bottom graph in Figure 1, represents the Irregular component which exhibits positive and negative jumps, being the last ones more frequent and of a greater magnitude. A compound Poisson process with jumps fits this idea and it is developed to improve the deterministic model.

After the previous analysis, the following model is proposed for the Ln prices:

$$P(t) = P_{TC}(t) + P_{EA}(t) + I(t),$$

where:

$$P_{TC}(t) = T(t) + C(t), \qquad \begin{cases} T(t) &= T(t) = b_0 + b_1 \cdot t, \\ C(t) &= b_2 \cdot \cos\left[\dfrac{2 \cdot \pi(t - b_3)}{468}\right], \end{cases} \tag{1}$$

being $P_{TC}(t)$ the Trend and Cyclical components, and $P_{EA}(t)$ and $I(t)$ the Seasonal and irregular components, respectively, which will be described later.

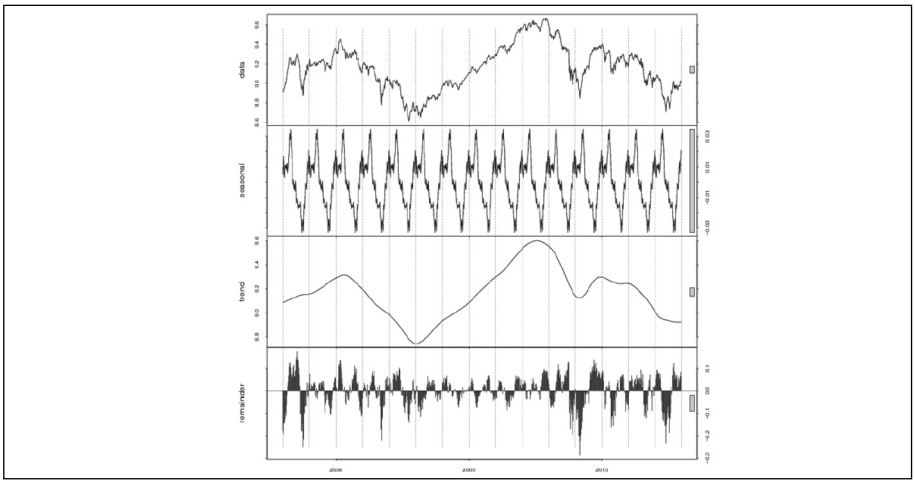


Figure 1. Graphical descomposion of the adjusted data in the components: Trend, Cyclical, Seasonal, and Irregular.

The Trend component is analyzed by a linear regression model on the Ln prices. The parameters of the Cyclical component have been obtained by the function *nls* in R (R Core Team, 2013). The results are shown in Table 2, being all the parameters statistically significantly.

	Coefficient	Std. Error	T-Value	P-Value
b_0	9.011	0.01002	899.23	0
b_1	0.0003943	0.00002232	17.67	0
b_2	0.2586	0.006832	37.85	0
b_3	43.24	2.103	20.56	0

Table 2. Parameters of the Trend and Cyclical model.

The next step is to remove the Trend and Cyclical components from the data.

The Annual Seasonal component is modeled by the trigonometric function (2). Table 3 shows the estimated model parameters.

$$P(t) - P_{TC}(t) = a_1 \cdot \cos\left[\frac{2 \cdot \pi(t - a_2)}{52}\right]. \tag{2}$$

	Coefficient	Std. Error	T-Value	P-Value
a_1	0.018353	0.006741	2.723	0.006621
a_2	10.465435	3.039692	3.443	0.000606

Table 3. Parameters of the Seasonal model.

Summarizing, the deterministic part of the model is as follows:

$$P(t) = 9.011 + 0.0004 \cdot t + 0.2586 \cdot \cos\left[\frac{2 \cdot \pi(t - 46.24)}{468}\right] + 0.0184 \cdot \cos\left[\frac{2 \cdot \pi(t - 10.4654)}{52}\right].$$

The Irregular component has been modeled with the Box-Jenkins (1973) methodology for time series. The analysis is implemented in R by *forecast* (2013). This tool determines the most appropriate ARIMA model according to the AIC-Akaike's Information Criterion (1974). The ARIMA (1,0,2) model is obtained from the data.

Several contrasts on the parameters and errors have been conducted to validate the model statistically: (1) the parameters of the model are statistically significant; (2) tests on the errors have been run to verify that: the mean is zero; the variance is constant; and to check the uncorrelation of the errors using the *tsdiag* function (Trapletti & Hornik, 2012) and, to test if the errors are normally distributed.

There are a number of extreme observations that are going to be referred as jumps. In the next step, the jumps are going to be isolated and modeled. The methodology applied in the analysis is described in Benth & Benth (2013). The jumps are isolated and it is assumed that the filtered residuals obtained are normally distributed. The p-value for the Kolmogorov-Smirnov test is 0.2296, which is greater than 0.05. Next, the jumps are modeled obtaining its size and frequency. Table 4 shows the descriptive values for the negative and positive jumps, as well as for altogether jumps. Altogether jumps are defined as the observed jumps in absolute value and the observations without jumps with null value.

	Positive Jumps	Negative Jumps	Altogether Jumps
Num. Observations	24	37	780
Mean	0.0742	-0.0835	0.0062
Std. Deviation	0.0203	0.0350	0.0230

Table 4. Values of ARIMA (1,0,2) parameters.

The proposed model comprises the sum of two compound Poisson processes, one for the positive jumps and the other for the negative ones, $s(t) = \mathrm{s}(t)^{+} + \mathrm{s}(t)^{-}$, being

$$S^{\pm}(t) = \sum_{i=1}^{N^{\pm}(t)} J_i^{\pm}.$$

The Poisson parameters are:

$$\lambda+ = 0.03076923, \qquad \lambda\text{-} = 0.0474359$$

They have been computed by dividing the number of positive and negatives jumps over the total number of observations. The length of the jump, estimated by exponential distribution with the mean of the altogether jumps, is 0.0062.

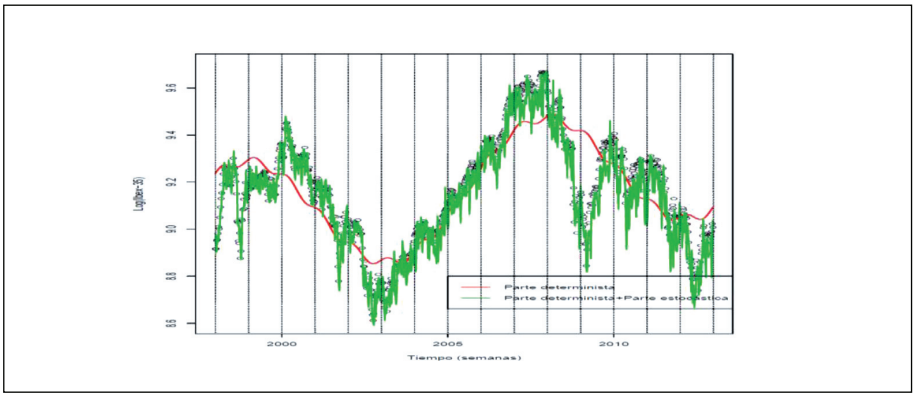


Figure 2. Complete model for the IBEX-35. The deterministic component, and the sum of the deterministic and the stochastic components.

Figure 2 shows the result of adjusting the complete model for the IBEX-35 disaggregated into the deterministic component, and the sum of the deterministic and the stochastic components.

To validate the model, a normality test on the final residuals has been conducted. The Kolmogorov-Smirnov test has obtained a p-value of 0.0562 (>0.05). Thus, they can be considered normally distributed. Finally, the model was built to forecast the prices of the IBEX-35 in the following nine months. Figure 3 shows the predictions for the data of the model, as well as, a 95% confidence interval.

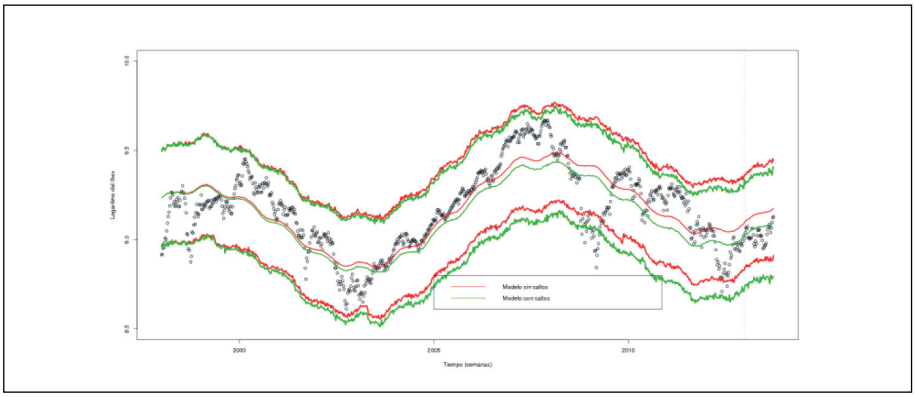


Figure 3. Complete model adjusted to the Ln IBEX-35 and nine months forecast using Monte Carlo simulations.

This interval has been obtained using the percentile method applied to 1,000 forecasted values for every week after simulating the random part of the model.

4. CONCLUSION

This paper proposes a model with two constituents: (1) a Deterministic component, with a nine-year cycle and one-year seasonal pattern, (2) an Irregular component, ARIMA (1,0,2) and a compound Poisson process with jumps. The analysis of the Irregular component is the main contribution of the study since it provides a different approach from the extant literature. Lastly, the model has been successfully validated; consequently, it qualifies for obtaining future predictions of the IBEX-35.

Therefore, the model can be applied to predict the dynamics of the IBEX- 35. The behavior of this index can be used to reduce the risk investment for experts and investors of the Spanish stock market.

ACKNOWLEDGEMENTS

This work was funded by grants from Ministry of Economy and Competitiveness (MTM2013-45381-P and MTM2013-41765-P).

REFERENCES

Akaike, H. (1974, Dec). "A New Look at the Statistical Model Identification". *Automatic Control, IEEE Transactions, 19* (6), 716-723.

Benth, F.E., and Benth, J.S. (2013). " Modeling and Pricing in Financial Markets for Weather Derivatives". *World Scientific*, 17.

Box, G., and Jenkins, G. (1973). "Some Comments on a Paper by Chatfield and Prothero and on a Review by Kendall". *Journal of the Royal Statistical Society*. Series A (General), 136 (3), 337-352.

Hyndman, R.J. (2013). *Forecast: Forecasting Functions for Time Series and Linear Models*. Retrieved from R package version 4.8: http://cran.r-project.org/web/packages/forecast/index.html

R Core Team (2013). "R: A Language and Environment for Statistical computing". *R Foundation for Statistical Computing*. Vienna, Austria: http://www.R-project.org/.

Trapletti, A., and Hornik, K. (2012). *R Package Time Series Analysis and Computational Finan*ce. Free software.

31. MEASURING THE SYSTEMIC RISK BY MULTIVARIATE INDICATORS

E. Di Bernardino
CNAM, Laboratoire Cédric EA4629, Département IMATH, France,
elena.di_bernardino@cnam.fr

J.M. Fernández-Ponce
Universidad de Sevilla, Departamento de Estadística e Investigación Operativa,
España, ferpon@us.es

F. Palacios-Rodríguez
Universidad de Sevilla, Departamento de Estadística e Investigación Operativa,
España, fpalacios2@us.es

M.R. Rodríguez-Griñolo
Universidad Pablo de Olavide, Departamento de Economía,
Métodos Cuantitativos e Historia Económica, España,
mrrodgri@upo.es

ABSTRACT

CoVaR (Conditional Value-at-Risk) is a risk measure proposed by Adrian and Brunnermeier (2011) which is able to quantify the systemic risk in the financial system. CoVaR indicates the Value-at-Risk for a financial institution that is conditional on a certain scenario. In this work, two alternative extensions of the classic univariate CoVaR are proposed in the multidimensional case. The level sets of multivariate distribution functions (resp. of multivariate survival distribution functions) are used to establish the two presented multivariate CoVaRs. Applications of the multivariate CoVaRs in the class of Archimedean copulas are given. Estimators for the multivariate proposed CoVaRs are illustrated in simulated studies and insurance real data.

1. INTRODUCTION

A risk-based methodology is gaining ground to supervise and regular the financial sector in both developing and industrialized countries. The need of regulators of quantifying market risk is part of this methodology. In spite of the fact that this problem has been traditionally handled in a univariate version by using Value-at-Risk (VaR), many multidimensional extensions of risk measures have been investigated in the past decade.

Jouini *et al*. (2004) introduce a class of set-value coherent risk measures. Embrechts and Puccetti (2006), Nappo and Spizzichino (2009), and Prékopa (2012) use the notion of quantile curve in order to generalize the univariate VaR. Moreover, Cousin and Di Bernardino (2013) and Cousin and Di Bernardino (2014) propose two multivariate extensions of the classic univariate Value-at-Risk and of the classic univariate Conditional-Tail-Expectation respectively.

On the other hand, although the risk allocation problem is only affected by internal risk, the solvability of the financial institutions could involve external risk, such as the systemic risk. CoVaR (Conditional Value-at-Risk) is a systemic risk measure proposed by Adrian and Brunnermeier (2011) which measures a financial institution's contribution to systemic risk and its contribution to the risk of other financial institutions. In its original univariate version, CoVaR indicates the Value at Risk for a financial institution that is conditional on a certain scenario. Adrian and Brunnermeier (2011) also define other systemic risk measure, ΔCoVaR, which involves VaR and CoVaR measures.

In this work, two alternative extensions of the classic univariate CoVaR based on the multivariate quantile settings of Embrechts and Puccetti (2006), Cousin and Di Bernardino (2013), and Cousin and Di Bernardino (2014) are proposed (Section 2). Furthermore, illustrations and estimations of these new risk measures are given under Archimedean copula conditions (Sections 3 and 4).

2. DEFINITIONS

In the following, consider that $\mathbf{X} = (X_1, \ldots, X_d)$ is a non-negative absolutelycontinuous random vector (with respect to Lebesgue measure λ on $\mathbb{R}^d$) with distribution function F and survival function $\overline{F}$. Moreover, the multivariate distribution function F is assumed to be partially strictly-increasing[1] such that $E(X_i) < \infty$ for $i = 1, \ldots, d$. Under these conditions, F is said to verify the regularity conditions.

Definition 2.1 (Multivariate Lower-Orthant CoVaR). Consider a random vector $\mathbf{X}$ which satisfies the regularity conditions. For $\alpha \in (0, 1)$, we define the multivariate lower-orthant CoVaR at probability level α by

$$\underline{\mathrm{CoVaR}}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \mathrm{VaR}_{\boldsymbol{\omega}}(\mathbf{X}|\mathbf{X} \in \partial\underline{L}(\alpha)) = \begin{pmatrix} \mathrm{VaR}_{\omega_1}(X_1|\mathbf{X} \in \partial\underline{L}(\alpha)) \\ \vdots \\ \mathrm{VaR}_{\omega_d}(X_d|\mathbf{X} \in \partial\underline{L}(\alpha)) \end{pmatrix},$$

where $\omega = (\omega_1, \ldots, \omega_d)$ is a marginal risk vector with $\omega_i \in [0, 1]$, for $i = 1, \ldots, d$, and $\partial\underline{L}(\alpha)$ is the boundary of the set $\underline{L}(\alpha) := \{\boldsymbol{x} \in \mathbb{R}^d_+ : F(\boldsymbol{x}) \geq \alpha\}$. Therefore, under regularity conditions,

$$\underline{\mathrm{CoVaR}}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \begin{pmatrix} \mathrm{VaR}_{\omega_1}(X_1|F(\mathbf{X}) = \alpha) \\ \vdots \\ \mathrm{VaR}_{\omega_d}(X_d|F(\mathbf{X}) = \alpha) \end{pmatrix}.$$

Similarly, the multivariate upper-orthant CoVaR can be defined.

Definition 2.2 (Multivariate Upper-Orthant CoVaR). Consider a random vector $\mathbf{X}$ which satisfies the regularity conditions. For $\alpha \in (0, 1)$, we define the multivariate upper-orthant CoVaR at probability level α by

[1] A function $F(x_1, \ldots, x_n)$ is partially strictly-increasing on $\mathbb{R}^d_+ \setminus \mathbf{0}$ if the function of one variable $g(\cdot) = F(x_1, \ldots, x_{j-1}, \cdot, x_{j+1}, \ldots, x_d)$ are strictly-increasing.

$$\overline{\mathrm{CoVaR}}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \mathrm{VaR}_{\boldsymbol{\omega}}(\mathbf{X}|\mathbf{X} \in \partial\overline{L}(\alpha)) = \begin{pmatrix} \mathrm{VaR}_{\omega_1}(X_1|\mathbf{X} \in \partial\overline{L}(\alpha)) \\ \vdots \\ \mathrm{VaR}_{\omega_d}(X_d|\mathbf{X} \in \partial\overline{L}(\alpha)) \end{pmatrix},$$

where $\omega = (\omega_1, \ldots, \omega_d)$ is a marginal risk vector with $\omega_i \in [0, 1]$, for $i = 1, \ldots, d$, and $\partial\underline{L}(\alpha)$ is the boundary of the set $\overline{L}(\alpha) := \{\boldsymbol{x} \in \mathbb{R}_+^d : \overline{F}(\boldsymbol{x}) \leq 1 - \alpha\}$. Therefore, under regularity conditions,

$$\overline{\mathrm{CoVaR}}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \begin{pmatrix} \mathrm{VaR}_{\omega_1}(X_1|\overline{F}(\mathbf{X}) = 1 - \alpha) \\ \vdots \\ \mathrm{VaR}_{\omega_d}(X_d|\overline{F}(\mathbf{X}) = 1 - \alpha) \end{pmatrix}.$$

Interesting properties of the above multivariate risk measures have been obtained. Particularly, the invariance and comonotonic dependence properties are studied.

3. APPLICATIONS OF MULTIVARIATE COVARS IN ARCHIMEDEAN COPULAS

Analytical expressions for multivariate CoVaR measures when X follows an Archimedean copula are obtained in this section. McNeil and Nešlehová (2009) introduced an important stochastic representation of Archimedean copulas (see Proposition 2.3 in Cousin and Di Bernardino, 2013). The next two characterizations by assuming Archimedean copula for the multivariate CoVaRs come from both the definition of Archimedean copula and the stochastic representation in McNeil and Nešlehová (2009).

Corollary 3.1. Let $\mathbf{X}$ be a d-dimensional random vector with an Archimedean copula with generator ϕ and $\alpha \in (0, 1)$. Therefore,

$$\underline{\mathrm{CoVaR}}^i_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \mathrm{VaR}_{\omega_i}\left[F^{-1}_{X_i}(\phi^{-1}(S_i\phi(\alpha)))\right], \ \textit{for } i = 1, \ldots, d, \tag{1}$$

where $\omega \in [0, 1]^d$ and S_i is a random variable with $Beta(1, d-1)$ distribution.

Corollary 3.2. Let **X** be a d-dimensional random vector with an Archimedean survival copula with generator ϕ and $\alpha \in (0, 1)$. Therefore,

$$\overline{\mathrm{CoVaR}}^{i}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \mathrm{VaR}_{\omega_i}\left[\overline{F}^{-1}_{X_i}(\phi^{-1}(S_i\phi(1-\alpha)))\right] \quad \textit{for } i = 1, \ldots, d, \tag{2}$$

where $\omega \in [0, 1]^d$ and S_i is a random variable with $Beta(1, d-1)$ distribution.

4. ESTIMATORS FOR MULTIVARIATE COVAR

An estimation procedure for the multivariate proposed CoVaR when **X** follows an Archimedean copula is shown in this section. Furthermore, the obtained estimators are illustrated with simulated and insurance real data. At first, let consider that **X** has an Archimedean copula structure. From Equation (1), a semiparametric estimator for the multivariate lower CoVaR is given by using a maximum pseudo-likelihood estimator of the dependence parameter θ associated with the generator of the copula and the empirical quantile estimation.

Definition 4.1. Let **X** be a d–dimensional random vector with Archimedean copula with generator ϕ_θ and $\alpha \in (0, 1)$. A semiparametric estimator of the i–component of the multivariate lower CoVaR is defined as

$$\widehat{\underline{\mathrm{CoVaR}}}^{i}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \widehat{\mathrm{VaR}}_{\omega_i}\left[\hat{F}^{-1}_{X_i}(\phi^{-1}_{\hat{\theta}_n}(S_i\phi_{\hat{\theta}_n}(\alpha)))\right], \quad \textit{for } i = 1, \ldots, d,$$

where $\omega \in [0, 1]^d$, S_i is a random variable with $Beta(1, d-1)$ distribution $\widehat{\mathrm{VaR}}_\omega(X)$ is the empirical estimator of $\mathrm{VaR}_\omega(X)$, $\phi_{\hat{\theta}_n}$ is the semiparametric estimator of ϕ_θ and $\hat{F}^{-1}_{X_i}$ is the empirical estimator of $\overline{F}^{-1}_{X_i}$ for $i = 1, \ldots, d$.

Secondly, let consider that X has an Archimedean survival copula structure. A semiparametric estimator of multivariate upper CoVaR considering the semiparametric estimator of the generator of the Archimedean survival copula and the empirical estimation of the quantile functions is introduced in Definition 4.2. This estimation comes from Equation (2).

Definition 4.2. Let **X** be a d–dimensional random vector with Archimedean survival copula with generator ϕ_θ and $\alpha \in (0, 1)$. A semiparametric estimator of the i–component of the multivariate upper CoVaR is defined as

$$\widehat{\overline{\mathrm{CoVaR}}}^{\,i}_{\alpha,\boldsymbol{\omega}}(\mathbf{X}) = \widehat{\mathrm{VaR}}_{\omega_i}\left[\hat{\overline{F}}^{-1}_{X_i}(\phi^{-1}_{\hat{\theta}_n}(S_i\phi_{\hat{\theta}_n}(1-\alpha)))\right], \qquad \textit{for } i = 1, \ldots, d,$$

where $\omega \in [0, 1]^d$, S_i is a random variable with $Beta(1, d-1)$ distribution $\widehat{\mathrm{VaR}}_\omega(X)$ is the empirical estimator of $\mathrm{VaR}_\omega(X)$, $\phi_{\hat{\theta}_n}$ is the semiparametric estimator of ϕ_θ and $\hat{F}^{-1}_{X_i}$ is the empirical estimator of $\overline{F}^{-1}_{X_i}$ for $i = 1, \ldots, d$.

The performance of the previous introduced estimators is evaluated in simulated studies. We focus on Definition 4.1 (the estimator proposed in Definition 4.2 could be similarly studied) and we restrict ourselves to the bivariate case (the illustrations could be extended to any dimension). For this purpose, the ratio $\widehat{\underline{\mathrm{CoVaR}}}^{1}_{\alpha,\boldsymbol{\omega}}(X,Y)/\underline{\mathrm{CoVaR}}^{1}_{\alpha,\boldsymbol{\omega}}(X,Y)$ is considered for two models, different values of α and ω, and two different sample sizes. For instance, the above ratio for a random vector (X, Y) which follows a bivariate Gumbel copula with parameter $\theta = 2$ and Pareto marginals with location parameter 1 and shape parameter 2 for $n = 1000$ is gathered in Figure 1. The aim is to analyse misspecification model error to study the bias and the variance of the estimation when the parametric form of the copula is not appropriate to the data. Moreover, the empirical standard deviation and the relative mean square error of the estimator of the multivariate lower CoVaR are studied.

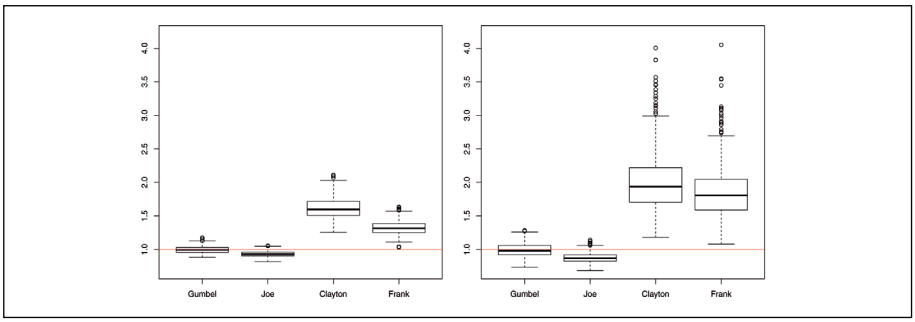


Figure 1. (X, Y) follows a bivariate Gumbel copula with parameter $\theta = 2$ and Pareto marginals with location parameter 1 and shape parameter 2. Box 1 plot for the ratio $\widehat{\underline{\mathrm{CoVaR}}}^{1}_{\alpha,\omega}/\underline{\mathrm{CoVaR}}^{1}_{\alpha,\omega}$ for $n = 1000$ with $\alpha = 0.75$ and $\omega = 0.9$ (left panel); $\alpha = 0.9$ and $\omega = 0.95$ (right panel).

Loss-ALAE data is an insurance real data set with n = 1500 claims. Each claim consists of an indemnity payment (the loss, X) and an allocated loss adjustment expense (ALAE, Y). The proposed estimators of the multivariate CoVaRs measures are obtained for Loss-ALAE data (in the log scale). In Figure 2, Loss-ALAE data in the log scale, the respective semiparametric estimated α-level sets ($\partial\underline{L}(\alpha)$ and $\partial\overline{L}(\alpha)$), the univariate empirical quantiles of Loss-ALAE data, and the estimators of multivariate CoVaRs measures are gathered.

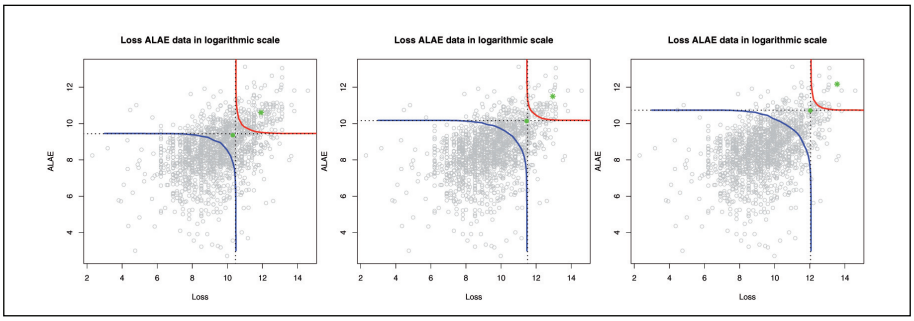


Figure 2. Loss ALAE data in log scale, boundary of estimated level sets ($\partial\underline{L}(\alpha)$, red line), boundary of estimated level sets ($\partial\overline{L}(\alpha)$, blue line), empirical quantile of Loss data (dotted black line), empirical quantile of ALAE data (dotted black line), $\widehat{\mathrm{CoVaR}}_{\alpha,\omega}$ (stars) and $\widehat{\overline{\mathrm{CoVaR}}}_{\alpha,\omega}$ (solid circles) with (α = 0.75, ω = 0.9) (left panel); (α = 0.9, ω = 0.95) (center panel); (α = 0.95, ω = 0.98) (right panel).

CONCLUSION

Two multivariate extensions of the classic CoVaR are presented for continuous random vectors in this work. The proposed measures are based on the corresponding quantile functions. Therefore, these measures are more robust to extreme values than any other central tendency measures. Analytical expressions for the multivariate CoVaRs measures are provided when random vectors follow an Archimedean copula. Furthermore, a estimation procedure for our multivariate risk measures is given in simulated data and insurance real data.

In a future perspective, quantile regression estimations by using extreme theory for the two multivariate CoVaRs can be analysed. Applications in financial sector of ΔCoVaR risk measures using the multivariate CoVaR proposed could be given.

ACKNOWLEDGMENTS

This work was partly supported by a grant from Junta de Andalucía (Spain) for research group (FQM- 328) and by a pre-doctoral contract (Palacios Rodríguez, F.) from the "V Plan Propio de Investigación" of the University of Seville.

REFERENCES

Adrian, T., and Brunnermeier, M.K. (2011). *CoVaR*. NBER Working Papers 17454, National Bureau of Economic Research, Inc.

Cousin, A., and Di Bernardino, E. (2013). "On Multivariate Extensions of Value-at-Risk". *Journal of Multivariate Analysis*, 119, 32-46.

Cousin, A., and Di Bernardino, E. (2014). "On Multivariate Extensions of Conditional-Tail-Expectation". *Insurance: Mathematics and Economics*, 55, 272-282.

Embrechts, P., and Puccetti, G. (2006). "Bounds for Functions of Multivariate Risks". *Journal of Multivariate Analysis*, 97 (2), 526-547.

Jouini, E., Meddeb, M., and Touzi, N. (2004). "Vector-Valued Coherent Risk Measures". *Finance Stoch.*, 8 (4), 531-552.

McNeil, A.J., and Nešlehová, J. (2009). "Multivariate Archimedean Copulas, d-Monotone Functions and *l*1-norm Symmetric Distributions". *The Annals of Statistics*, 37, 3059-3097.

Nappo, G., and Spizzichino, F. (2009). "Kendall Distributions and level Sets in Bivariate Exchangeable Survival Models". *Information Sciences*, 179, 2878-2890.

Prékopa, A. (2012). "Multivariate Value at Risk and Related Topics". *Annals of Operation Research*, 193, 49-69.

32. ESTIMATION OF MULTIVARIATE CRITICAL LAYERS: APPLICATIONS TO RAINFALL DATA

E. Di Bernardino
CNAM, Paris, Département IMATH, France,
elena.di_bernardino@cnam.fr

D. Rullière
Université de Lyon 1, ISFA, Laboratoire SAF, France,
didier.rulliere@univ-lyon1.fr

ABSTRACT

Calculating return periods and critical layers (i.e. multivariate quantile curves) in a multivariate environment is a difficult problem. A possible consistent theoretical framework for the calculation of the return period, in a multi-dimensional environment, is essentially based on the notion of copula and level sets of the multivariate probability distribution. In this paper we propose a fast and parametric methodology to estimate the multivariate critical layers of a distribution and its associated return periods. The model is based on transformations of the marginal distributions and transformations of the dependence structure within the class of Archimedean copulas. The methodology is illustrated on rainfall 5-dimensional real data. We also develop a nested model on this rainfall 5-dimensional real data set.

1. INTRODUCTION

The notion of Return Period (RP) is frequently used in environmental sciences for the identification of dangerous events, and provides a means for rational decision making and risk assessment. Roughly speaking, the RP can be considered as an analogue of the "Value-at-Risk" in Economics and Finance, since it is used to quantify and assess the risk (see, e.g., Nappo and Spizzichino, 2009). In engineering

practice, finance, insurance and environmental science the choice of the RP depends on the impact/magnitude of the considered event and the consequences of its realisation.

Equally important is the related concept of *design quantile*, usually defined as "the value of the variable characterizing the event associated with a given RP". In the univariate case the design quantile is usually identified without ambiguity. Conversely in the multivariate setting different definitions are possible. For this reason, the identification problem of design events in a multivariate context has recently attracted the attention of many researches (see for instance Salvadori *et al.*, 2007).

In the following, we will consider a sequence $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ of independent and identically distributed d–dimensional random vectors, with $d > 1$.

Thus each $\mathbf{X}_k$, $k \in \mathbb{N}$, has the same multivariate distribution $F_X : \mathbb{R}^d_+ \rightarrow [0, 1]$ as the nonnegative real-valued random vector $\mathbf{X} \sim F_X = C(F_{X1}, \ldots, F_{Xd})$ describing the hydrological phenomenon under investigation. The function C is the d-dimensional copula associated to F. We write $I = \{1, \ldots, d\}$ the set of indexes of the considered random variables and of their associated cumulative distribution functions, i.e., $Fx_i(x_i) = P(X_i \leq x_i)$, for $i \in I$.

Definition 1.1 (Critical layer). The critical layer $\partial L(\alpha)$ associated to the multivariate distribution function F_X of level $\alpha \in (0, 1)$ is defined as

$$\partial L(\alpha) = \{\boldsymbol{x} \in \mathbb{R}^d : F_{\boldsymbol{X}}(\boldsymbol{x}) = \alpha\}.$$

Then $\partial L(\alpha)$ is the iso-hyper-surface (with dimension d – 1) where F equals the constant value α. The critical layer $\partial L(\alpha)$ partitions $\mathbb{R}^d$ into three non-overlapping and exhaustive regions:

$$\begin{cases} L^<(\alpha) &= \{\mathbf{x} \in \mathbb{R}^d : F_{\mathbf{X}}(\mathbf{x}) < \alpha\}, \\ \partial L(\alpha) &= \text{the critical layer itself}, \\ L^>(\alpha) &= \{\mathbf{x} \in \mathbb{R}^d : F_{\mathbf{X}}(\mathbf{x}) > \alpha\}. \end{cases}$$

The Return Period is defined as the average time required for reaching the set $L^{>}(\alpha)$, that is:

$$\mathrm{RP}^{>}(\alpha) = \Delta_t \cdot \mathbb{E}[N] = \frac{\Delta_t}{\mathbb{P}[\mathbf{X} \in L^{>}(\alpha)]}, \tag{1}$$

where $\Delta_t > 0$ is the (deterministic) average time elapsing between X_k and X_{k+1}, $k \in \mathbb{N}$. The probability that a realization of this vector belongs to $L^{<}(\alpha)$ is given by the Kendall's function, which only depends on the copula C of this random vector, i.e.,

$$K_C(\alpha) = \mathbb{P}[X \in L^{<}(\alpha)] = \mathbb{P}[C(U_1, \ldots, U_d) \leq \alpha], \quad \text{for } \alpha \in (0,1). \tag{2}$$

Then, the considered Return Period can be expressed using Kendall's function in (2), $\mathrm{RP}^{>}(\alpha) = \Delta_t \cdot \frac{1}{1-K_C(\alpha)}$. .

This paper aims at:

- giving a parametric representation of the multivariate distribution F of a random vector X, here representing rain measurements,
- giving direct estimation procedure for this representation,
- giving closed parametric expressions, both for critical layers in Definition 1.1 and Return Periods in (1),
- adapting this methodology to some asymmetric dependencies (as, for instance, non-exchangeable random vectors) by using nested model for the recording rainfall stations.

In the next section, we introduce the model used to answer the issues introduced above.

2. THE TRANSFORMED MODEL

We consider the following model, which is detailed in Di Bernardino and Rullière (2013a),

$$\widetilde{F}(x_1,\ldots,x_d) = T \circ C_0(T_1^{-1} \circ F_1(x_1),\ldots,T_d^{-1} \circ F_d(x_d)),$$

or equivalently

$$\left\{ \begin{array}{lcl} \widetilde{F}(x_1,\ldots,x_d) & = & \widetilde{C}(\widetilde{F}_1(x_1),\ldots,\widetilde{F}_d(x_d)), \text{ with} \\ \widetilde{C}(u_1,\ldots,u_d) & = & T \circ C_0(T^{-1}(u_1),\ldots,T^{-1}(u_1)) \\ \widetilde{F}_i(x) & = & T \circ T_i^{-1} \circ F_i(x), \text{ for } i \in I, \end{array} \right.$$

where $F_1, \ldots, F_d$ are given parametric initial marginal cumulative distribution functions, and where C_0 is a given initial copula. Hence the distribution $\widetilde{F}(x_1,\ldots,x_d)$ is built from transformed marginals $\widetilde{F}_i, i \in I$ and from a transformed copula $\widetilde{C}$, under regularity conditions. Transformation T permits to transform the initial dependence structure C_0. For a given T, transformations T_i permit to transform marginals, $i \in I$. Furthermore we will assume in the following that C_0 is an Archimedean copula. In our paper we show how to estimate the transformations T and T_i, $i \in I$. Estimation procedure is omitted here for sake of brevity. For further details the interested reader is referred to Di Bernardino and Rullière (2014).

3. NUMERICAL RESULTS ON THE RAINFALL 5-DIMENSIONAL REAL DATA

In the following, we present some estimation results on rainfall 5-dimensional real data provided using our estimation procedure summarized in Section 2. Data comes from the website CISL Research Data Archive (RDA). Geographical position of 5 stations and the scatter plot of ranks of data are provided in Figure 1. We perform a Goodness-of-Fit test based on the empirical process in order to test the quality of the adjustment of our transformed copula $\widetilde{C}$ on these multivariate data. We obtain a p-value equal to 0.38. In the large scale Monte Carlo experiments carried out by Genest *et al.* (2009), the statistic S_n gave the best results overall. An approximate p-value for S_n can be obtained by means of a parametric bootstrap-based procedure (see results in Table 1). In the following we intend to show the flexibility of the proposed model and associated estimation procedure. In particular, we adapt our methodology in the case of some asymmetric dependencies (as, for instance, non-exchangeable random vectors). The correlation matrix of the considered rainfall data is displayed in Figure 2 (left).

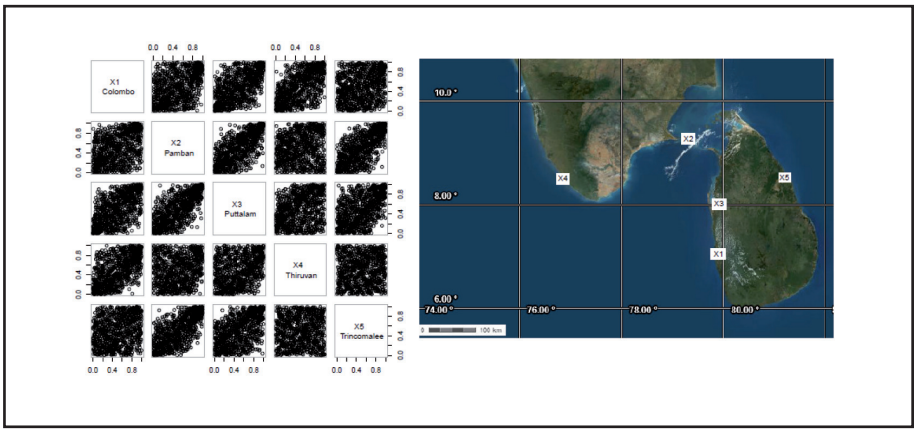


Figure 1. Left: Scatter plot of ranks for the considered 797 monthly rainfall measurements (in decimeter) in 5 stations of Sri-Lanka and India between January 1893 and June 2013. Right: Geographical positions of 5 considered stations.

Copula under H_0	S_n	S^B_n	S_C	A_n
Gumbel-Hougaard	0.00331	0.00495	0.00454	0.03465
Clayton	0.00381	0.00980	0.00704	0.00981
Frank	0.00617	0.00941	0.00819	0.08416
t-Student	0.00495	0.00592	0.00498	0.00963
Normal	0.00980	0.00719	0.00454	0.00205
Joe	0.00819	0.00495	0.00454	0.00916

Table 1. The bootstrapped p-values for different Goodness-of-Fit tests (see Genest *et al.*, 2009) for competitor copula families on the considered 5-dimensional rainfall data, with n = 797. In all cases, the number of Monte Carlo experiments is fixed at $N = 1{,}000$.

As we can see, some pairs of stations present a bigger correlation. To illustrate how our model can be adapted to this situation we have decided to create 2 different clusters. An hierarchical cluster analysis on the set of dissimilarities produce by the distance of the X_i is developed. We obtain the result in Figure 2. As one can see, whatever the distance chosen for dissimilarities, the dendrogram gives a justification to chosen clusters of station indexes {2, 3, 5} and {1, 4}. Then, we firstly fit a 3-dimensional model for the first group and a 2-dimensional one for the second one. We generate the pseudo-data coming from these two models and finally we construct the joint (root) copula for these bivariate data-set. In the following we take as initial copula C_0 the independent one, and the initial margins $F_i(x) = 1 - e^{-x}$, $i \in A, B$.

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 \\ X_1 & \mathbf{1} & 0.377 & 0.552 & 0.504 & 0.207 \\ X_2 & 0.377 & \mathbf{1} & \mathbf{0.698} & 0.178 & \mathbf{0.660} \\ X_3 & 0.552 & \mathbf{0.698} & \mathbf{1} & 0.276 & 0.523 \\ X_4 & 0.504 & 0.178 & 0.276 & \mathbf{1} & 0.018 \\ X_5 & 0.207 & \mathbf{0.660} & 0.523 & 0.018 & \mathbf{1} \end{bmatrix}$$

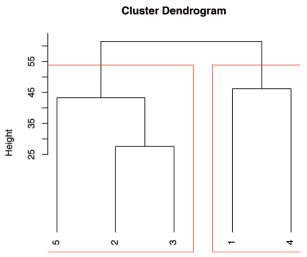


Figure 2. Left: Correlation matrix of the considered rainfall data. Correlations greater than 60% are indicated in bold font. Right: Dendrogram resulting to the hierarchical cluster analysis on the set of dissimilarities produced by the Euclidian distance on the rainfall data. Red boxes show the two considered clusters.

The multivariate distribution for the cluster A = {2, 3, 5} is assumed to be written:

$$\begin{aligned} F_A(\mathbf{x}_{2,3,5}) &= T_A \circ C_0(T_A^{-1} \circ \widetilde{F}_2(x_2),\, T_A^{-1} \circ \widetilde{F}_3(x_3),\, T_A^{-1} \circ \widetilde{F}_5(x_5)), \\ \text{with } \tilde{F}_i &= T_A \circ T_{A_i}^{-1} \circ F_i\,, \quad \text{for } i \in A. \end{aligned} \tag{3}$$

The multivariate distribution for the cluster B = {1, 4} is assumed to be written:

$$\begin{aligned} F_B(\mathbf{x}_{1,4}) &= T_B \circ C_0(T_B^{-1} \circ \widetilde{F}_1(x_1),\, T_B^{-1} \circ \widetilde{F}_4(x_4)), \\ \text{with } \tilde{F}_i &= T_B \circ T_{B_i}^{-1} \circ F_i\,, \quad \text{for } i \in B. \end{aligned} \tag{4}$$

The whole 5–dimensional distribution is assumed to be written:

$$\widetilde{F}(\mathbf{x}_{1,2,3,4,5}) = T \circ C_0\left(T^{-1} \circ F_A(\mathbf{x}_{2,3,5}),\, T^{-1} \circ F_B(\mathbf{x}_{1,4})\right), \tag{5}$$

where $\tilde{C}(u, v) = T \circ C_0\,(T^{-1}(u), T^{-1}(v))$ is referred as the root copula at point *(u, v)*. It is effectively a proper copula if the transformation T satisfies admissibility conditions that are given in Proposition 2.1 of Di Bernardino and Rullière (2013b).

To estimate the external transformation T of model (5) we firstly construct a bivariate pseudo data-set:

$$Z_1 = F_A(X_1, X_4), \quad Z_2 = F_B(X_2, X_3, X_5).$$

Then we fit on this bivariate data-set a model

$$\begin{aligned} \widetilde{F}_{(Z_1,Z_2)}(z_1,z_2) &= T \circ C_0(T^{-1} \circ \tilde{F}_1(z_1),\, T^{-1} \circ \tilde{F}_2(z_2)), \\ \text{with } \tilde{F}_i &= T \circ T_1^{-1} \circ F_i\,, \quad \text{for } i = 1, 2. \end{aligned}$$

Let $\alpha \in (0, 1)$ be a targeted level for a critical layer. Let C_0 be the initial copula to be transformed, and assume that C_0 is the independent copula. Then the analytical critical layers of the distributions $\widetilde{F}_{1,4}$ and $\widetilde{F}_{2,3,5}$ are easy to obtain. For $\widetilde{F}_{1,4}$ in (4), we have

$$\partial L_B(\alpha) = \left\{(x_1, x_4) : x_1 = \widetilde{F}_1^{-1} \circ T_{1,4}\left((T_{1,4}^{-1}(\alpha))^p\right), x_4 = \widetilde{F}_4^{-1} \circ T_{1,4}\left((T_{1,4}^{-1}(\alpha))^{1-p}\right)\right\}.$$

Analytical expressions of the inverse of any transformed margins $\widetilde{F}_i = T \circ T_i^{-1} \circ F_i$, for $i \in B$, are available since inverse transformations are given and since the initial distribution Fi is chosen to be readily invertible. Analogously, we get, for F_A in (3)

$$\begin{aligned} \partial L_A(\alpha) = \{(\mathbf{x}_{2,3,5}) : x_2 = \widetilde{F}_1^{-1} \circ T_{1,4}\left((T_{1,4}^{-1}(\alpha))^{p_1}\right), x_3 = \widetilde{F}_3^{-1} \circ T_{1,4}\left((T_{1,4}^{-1}(\alpha))^{p_2}\right) \\ x_5 = \widetilde{F}_5^{-1} \circ T_{1,4}\left((T_{1,4}^{-1}(\alpha))^{1-p_1-p_2}\right),\ p_1, p_2 \in (0,1),\ p_1 + p_2 < 1\}. \end{aligned}$$

For the nested distribution $\widetilde{F}$ in (5), one can write,

$$\begin{aligned} \partial L(\alpha) &= \left\{(x_1, \ldots, x_5) : \widetilde{F}(\mathbf{x}_{1,2,3,4,5}) = \alpha\right\} \\ &= \left\{(x_1, \ldots, x_5) : T \circ C_0\left(T^{-1} \circ \widetilde{F}_{2,3,5}(\mathbf{x}_{2,3,5}), T^{-1} \circ \widetilde{F}_{1,4}(\mathbf{x}_{1,4})\right) = \alpha\right\} \\ &= \left\{(x_1, \ldots, x_5) : T^{-1} \circ \widetilde{F}_{2,3,5}(\mathbf{x}_{2,3,5}) \cdot T^{-1} \circ \widetilde{F}_{1,4}(\mathbf{x}_{1,4}) = T^{-1}(\alpha)\right\} \end{aligned}$$

An illustration of critical-layers $\partial L_A(\alpha)$ and $\partial L_B(\alpha)$ derived above is provided in Figure 3.

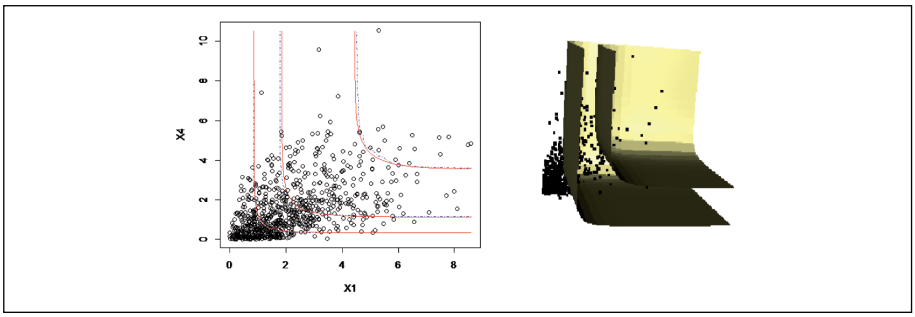


Figure 3. Left: 2-dimensional critical-layers $\partial L_B(\alpha)$ with α = 0.2, 0.5, 0.9 with associated empirical critical-layers (blue dashed lines). Right: 3-dimensional critical layers $\partial L_A(\alpha)$ with α = 0.3, 0.9. Black dots represent rainfall data (X_1, X_4) (left) and (X_2, X_3, X_5) (right).

REFERENCES

Di Bernardino, E., and Rullière, D. (2015). "Estimation of Multivariate Critical Layers: Applications to Hydrological Data (2015)". Accepted for publication in *Journal de la Société Française de Statistique*. Preprint available on HAL.

Di Bernardino, E., and Rullière, D. (2013a). "Distortions of Multivariate Distribution Functions and Associated Level Curves: Applications in Multivariate Risk Theory". *Insurance: Mathematics and Economics*, 53 (1): 190-205.

Di Bernardino, E., and Rullière, D. (2013b). "On Certain Transformations of Archimedean Copulas: Application to the Non-Parametric Estimation of their Generators". *Dependence Modeling*, 1: 1-36.

Genest, C., Rèmillard, B., and Beaudoin, D. (2009). "Goodness-of-Fit Tests for Copulas: A Review and a Power Study". *Insurance: Mathematics and Economics*, 44 (2): 199-213.

Nappo, G., and Spizzichino, F. (2009). "Kendall Distributions and Level Sets in Bivariate Exchangeable Survival Models". *Information Sciences*, 179: 2878-2890

Salvadori, G., De Michele, C., Kottegoda, N., and Rosso, R. (2007). *Extremes in Nature: An Approach Using Copulas*. Springer-Verlag: Berlin.

33. FINANCIAL TIME SERIES OBTAINED BY AGENT-BASED SIMULATION

P. Ferrando
Universitat de Barcelona, Spain,
pferrandohernandez@gmail.com

J. Fortiana
Universitat de Barcelona, Spain,
fortiana@ub.edu

ABSTRACT

We have designed and implemented a simple agent-based stock market simulating the evolution of a single stock price on the NetLogo platform. Agents represent investors buying and selling based on their individual decisions. The output series of returns is analyzed and shown to fit an ARCH model.

1. INTRODUCTION

Individual investors' decisions control the dynamics of stock markets. Agentbased simulation tries to reproduce this microstructure by creating software objects (agents) mimicking this behavior. The collective behavior of these agents, each of which is acting independently, produces price movements. In this paper, taking Silva's (2014) Collective Behavior in the Stock Market model as a departure point, we model the evolution of a very simple market, with a single asset price, using the NetLogo environment. Notwithstanding the naivety of the model, the simulated market has a remarkable realistic flavor: typically the simulated returns series can be fit to an ARCH model.

In Section 2, we describe the assumptions of the model about the different groups of investors considered, their general behavior in the market and the parameters used to model this behavior. In Section 3 we comment on some aspects of the implementation. In Section 4 we show an analysis of a representative output series of returns.

2. MODEL HYPOTHESES

Based on Silva's (2014) Collective Behavior in the Stock Market NetLogo model, we consider a market populated by 10.000 agents divided into two groups: investors affected by the disposition effect and investors who use stop-loss orders. Independently to the group they belong, all investors use long buying and short selling as strategies.

Agents collect information on whether their neighbors are buying (or selling) and will be willing to buy (or sell) depending on this information. In addition, at any moment an investor *j* may have additional information recommending to buy or sell. To model this situation, we use an auxiliary variable $O^{jt}_{buy(or\ sell)}$ for agent *j* and time step *t*. For example, if $O^{jt}_{buy} = 1$, the agent possesses extra information advising him to buy; if not, $O^{jt}_{buy} = 0$, and analogously for selling. We balance the influence of his preferred choice and his surrounding behavior. We use a weighting term ω_{jt} representing the investor's confidence in his own information: if $\omega_{jt} = 0$, the agent pays no attention to the information he owns; if $\omega_{jt} = 1$, his own information receives the same weight as that of his neighbors; when ω_{jt} grows, the weight given to his own information increases towards overconfidence. Finally, we assume that ω_{jt} follows a uniform distribution on the interval $[0, \omega_{max}]$ where ω_{max} is a maximum value previously established.

On the other hand, investors get information on the global sentiment of the market, so we define a market sentiment index I by:

$$I = D_t + I_{shock}$$

where

$$D_t = \frac{b_t - s_t}{n}, \tag{1}$$

$n = 10000$, b_t and st are the total number of buyers and sellers in the time period t, respectively, and I_{shock} is a constant related to additional information ($I_{shock} > 0$ is related to good news about the company, while $I_{shock} < 0$ to bad news and $I_{shock} = 0$ means that there is no important information). Nevertheless, each agent uses I differently, and the random parameter $\beta_j \in [0, 1)$ captures this fact. $\beta_j = 0$ means that agent j is not influenced by market sentiment, whereas $\beta_j = 1$ means that agent j keeps in mind this information to make decisions.

In summary, the probability that agent j buys stocks in a given time period t is:

$$P_{jt}^{buy} = \frac{b_{jt} + \omega_{jt} O_{jt}^{buy}}{8 + \omega_{jt} O_{jt}^{buy}} + \beta_j I \text{ , if } I > 0$$

where b_{jt} is the number of buyers around agent j (maximum 8) and the last term $(\beta_j I)$ is not considered if $I \leq 0$.

The probability that agent j sells stocks in a given time period t is:

$$P_{jt}^{sell} = \frac{s_{jt} + \omega_{jt} O_{jt}^{sell}}{8 + \omega_{jt} O_{jt}^{sell}} + \beta_j |I| \text{ , if } I < 0$$

where s_{jt} is the number of sellers around agent j (maximum 8) and the last term $(\beta_j |I|)$ is not considered if $I \geq 0$.

By default, heterogeneous expectations are considered, so every agent has a unique expected return. In addition, expected returns follow a uniform distribution on the interval $[0.01, r_{max})$, where r_{max} is a maximum value previously established.

Finally, we consider a sigmoidal (hyperbolic tangent) form for the excess demand, resulting in the following stock price in the time period t:

$$p_t = p_{t-1} (1 + \tanh D_t) \tag{2}$$

3. IMPLEMENTATION DETAILS

At each time period, agents have either a long position, or a short position or fetch information from their neighbors and the market sentiment to enter the market.

Investors with a long position close it depending on the price. If sale at the current price produces a revenue higher than their individual expected return, they sell. Closing a losing position depends on the type of investor. Investors using stop-loss orders sell if the price decrease results in a loss greater than their individual stop-loss boundary. Investors affected by the disposition effect sell if losses are greater than 2.25 times their expected return. Finally, if the price has not reached these limits, both types of investors hold their position in this time period.

Similarly, investors with a short position close it depending on the price. If buying at the current price produces a revenue higher than their individual expected return, they buy. Closing a losing position also depends on the type of investor. Investors using stop-loss orders buy if the price increase results in a loss greater than their stop-loss boundary. Investors affected by the disposition effect buy when losses are greater than 2.25 times their expected return. Finally, if the price has not reached these limits, both types of investors hold their position in this time period. If investors do not buy or sell, they enter the market with a probability $\mathrm{buyprob}$ of long buying and $\mathrm{sellprob}$ of short selling. Otherwise, agents do not operate.

4. ANALYSIS OF THE RETURNS SERIES

We show here a representative run of the simulation process, with n=10000 agents, expected return limit r_{max} = 0.15, ω_{max} = 1 and I_{shock} = 0 . Also, the ratio of investors affected by the disposition effect to investors using stoploss orders is 3. Conventionally, the initial price is set to 100, although it does not affect the model dynamics.

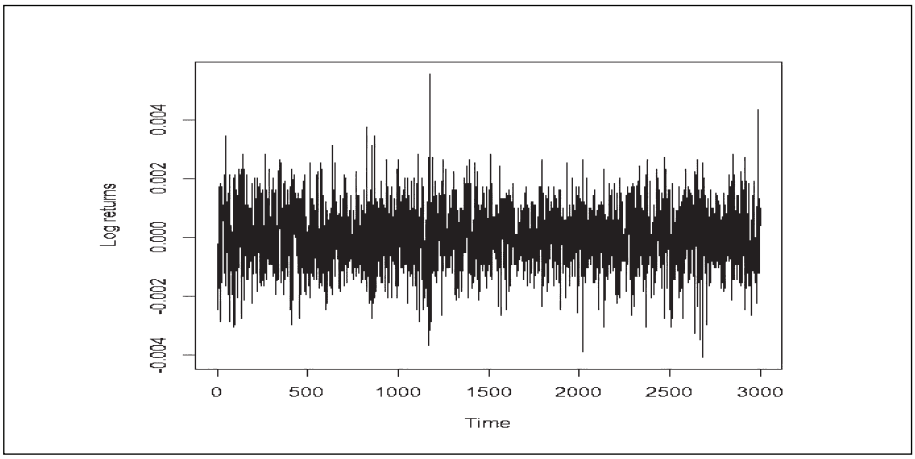


Figure 1. Time plot of 3,000 simulated log returns.

The distribution of returns is, as it could be expected in a realistic scenario, heavy tailed (see Figure 2).

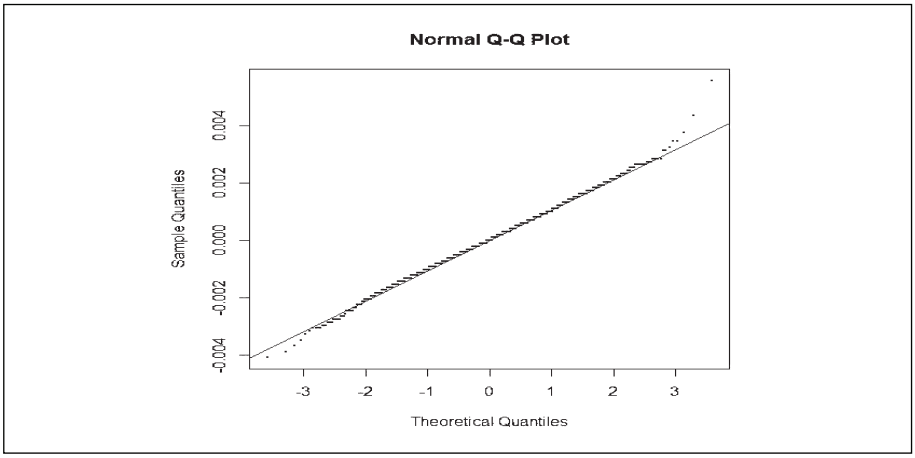


Figure 2. Q-Q plot of simulated log returns versus a normal distribution.

The series seems to be stationary (see Figure 1), and the observations are serially dependent according to the correlogram (Figure 3), so we use an ARMA model to fit the returns. Based on the AIC, we obtain an ARMA(1,2) given by:

$$r_t = 0.92\, r_{t-1} + z_t - 1.12\, z_{t-1} + 0.27\, z_{t-2} \tag{3}$$

where $\{z_t\}$ is assumed to be white noise.

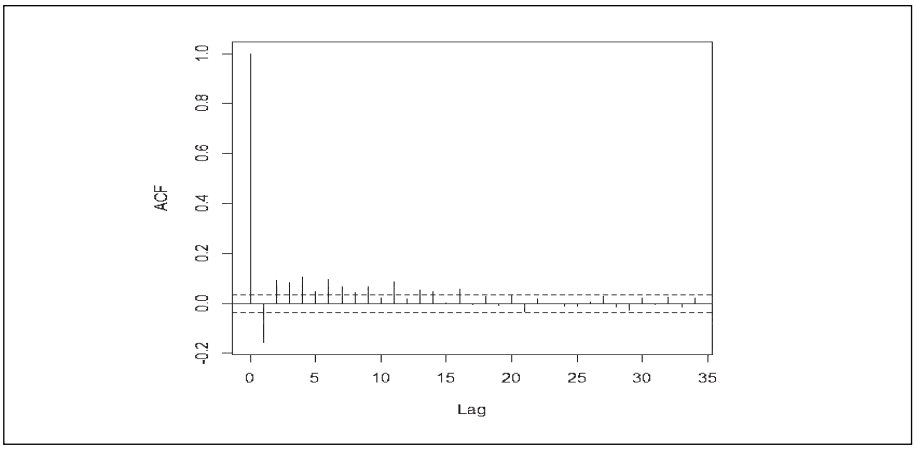


Figure 3. Correlogram of the simulated series.

However, observing the residuals, we realize that the ARMA fit solves the problem with the correlated observations but cannot explain the correlation in the variance (Figure 4). Therefore, we need an ARCH model to fit the residuals. Based on the AIC, we obtain an ARCH (3).

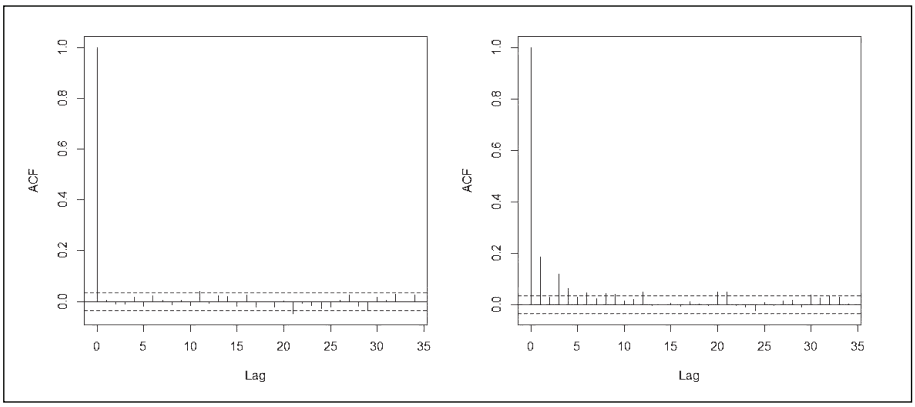


Figure 4. Correlogram of residuals (on the left) and squared residuals (on the right) of the ARMA model.

In summary, the complete model for the log returns is the ARMA (1,2)-ARCH(3) given by:

$$r_t = 0.92\, r_{t-1} + z_t - 1.12\, z_{t-1} + 0.27\, z_{t-2} \tag{4}$$

where

$$z_t = a_t \sqrt{7.38 \times 10^{-7} + 0.19\, z_{t-1}^2 + 0.09\, z_{t-3}^2}$$

and $\{a_t\}$ is white noise.

5. DISCUSSION

This work is an initial step in a study of the relation between agent properties, as described by their functionality and parameters, and the statistical behavior of the resulting series of returns.

ACKNOWLEDGMENTS

Special Thanks to Elder Silva for sharing many useful informations.

REFERENCES

Silva, E. (2014). *Collective Behavior in the Stock Market*. NetLogo Modeling Commons. http://modelingcommons.org

Tesfatsion, L., and Judd, K.L. (eds.) (2006), *Handbook of Computational Economics*. Vol. 2: *Agent-Based Computational Economics*. North-Holland/Elsevier, Amsterdam, Netherlands.

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, North-western University. Evanston, IL.

34. EULER-POISSON SCHEMES FOR LÉVY PROCESSES

A. Ferreiro-Castilla
Banc Sabadell, Spain,
aferreiro.cl@gmail.com

ABSTRACT

In this note we will contextualize the recently established Wiener-Hopf Monte Carlo (WHMC) simulation technique for Lévy processes from Kuznetsov *et al.* (2011) into a more general framework allowing us to use the same technique in a larger set of problems. We will briefly show how the scheme can be used to approximate Lévy driven SDEs or how to approximate different types of path dependent quantities. In a way, the present note summarizes and connects a set of results contained in Ferreiro-Castilla *et al.* (2014), (2015) and (2016), therefore we intentionally leave most of the technicalities aside for this note.

1. INTRODUCTION

Let $X := (X_t)_{t \geq 0}$ be a Lévy process, i.e. a (real valued) stochastic process starting from 0 with cadlag paths (right continuous and with left limits) and stationary, independent increments. The Lévy-Khintchine formula entails that the characteristic exponent Ψ, defined as $\mathbb{E}\left[e^{izX_t}\right] = e^{-t\Psi(z)}$ for all $t \geq 0$ and $z \in \mathbb{R}$, can be expressed as

$$\Psi(z) = \frac{\sigma^2}{2} z^2 + \mathrm{i}az + \int_{\mathbb{R}\setminus\{0\}} (1 - e^{\mathrm{i}zx} + \mathbf{1}_{\{|x|<1\}} \mathrm{i}zx) \Pi(\mathrm{d}x) \,, \tag{1}$$

where σ, $a \in \mathbb{R}$ and Π a measure on $\mathbb{R}\setminus\{0\}$ satisfying $\int_{\mathbb{R}\setminus\{0\}} (x^2 \wedge 1)\Pi(\mathrm{d}x) < \infty$. Of interest in several fields are quantities depending on the path of X, the complete path of X is numerically intractable and, ultimately, any numerical scheme can only be

based on simulating the increments of the driving process. A first approach consists in computing an Euler skeleton of the path; but even for a Brownian motion, where the Euler scheme is exact, computing a simple supremum in this way leads to a significant bias (see Broadie *et al.*, 1999) due to the fact that the skeleton misses the excursions of the process. In this note we are interested in constructing a skeleton for the path based on the Wiener-Hopf factorization. Suppose that for any $q > 0$, $\mathrm{e}(q)$ is an exponentially distributed random variable with mean q^{-1} that is independent of X. Recall that $\overline{X}_t = \sup_{s \le t} X_s$ and let $\underline{X}_t := \inf_{s \le t} X_s$. The Wiener-Hopf factorisation states that the random variables $\overline{X}_{\mathrm{e}(q)}$ and $\overline{X}_{\mathrm{e}(q)} - X_{\mathrm{e}(q)}$ are independent. Thanks to the so-called principle of duality, that is to say the equality in law of the pair $\{X_{(t-s)-} - X_t : 0 \le s \le t\}$ and $\{-X_s : 0 \le s \le t\}$, it follows that $\overline{X}_{\mathrm{e}(q)} - X_{\mathrm{e}(q)}$ is equal in distribution to $\underline{X}_{\mathrm{e}(q)}$. This leads to the following factorisation of characteristic functions

$$\mathbb{E}(\mathrm{e}^{\mathrm{i}\theta X_{\mathbf{e}(q)}}) = \mathbb{E}(\mathrm{e}^{\mathrm{i}\theta \overline{X}_{\mathbf{e}(q)}}) \times \mathbb{E}(\mathrm{e}^{\mathrm{i}\theta \underline{X}_{\mathbf{e}(q)}}), \qquad \text{for all } \theta \in \mathbb{R},$$

known as the Wiener-Hopf factorisation. Equivalently,

$$X_{\mathbf{e}(q)} \stackrel{d}{=} S_q + I_q, \tag{2}$$

where S_q and I_q are independent and equal in distribution to $\overline{X}_{\mathrm{e}(q)}$ and $\underline{X}_{\mathbf{e}(q)}$, respectively. Here we use the notation $\stackrel{d}{=}$ to mean equality in distribution.

The above factorization captures in one time step the nature of the path; i.e. knowing the supremum and the infimum one would be able to answer questions relating whether the process crossed a barrier or whether there was a big jump in the time step. Obviously the time step is itself an random variable, but thanks to the equality

$$\sum_{i=1}^{n} \frac{1}{n} \mathbf{e}_i(1) \stackrel{a.s.}{\to} 1 \text{ as } n \uparrow \infty \tag{3}$$

we can create a random partition of the interval [0, 1] with a mesh size $\mathbb{E}[\frac{1}{n}\mathbf{e}_i(1)] \to 0$ as $n \uparrow \infty$. We call our method the Euler-Poisson scheme, as one can think of the random partition as given by the arrival times of a Poisson process.

For the sake of simplicity, recall that $\sum_{i=1}^{k} \frac{1}{n} e_i(1)$ has the same distribution as a Gamma random variable which will be denoted by *g(k, n)*. Therefore the Euler-Poisson scheme of the process X refers to the skeleton $\{X_{ek(1/n)}\}_{k\geq 0}$ or $\{X_{g(k,n)}\}_{k\geq 0}$.

1.1. Heuristics behind the scheme

Let us give a brief discussion of the heuristics of the Euler-Poisson scheme and an idea why this scheme should be a particular good approach to compute path dependent quantities of Lévy processes.

It can be inferred from Doney (2004) that, for all $k > 0$, the random variables

$$M_k := \sup_{g(k,n)\leq t<g(k+1,n)} X_t \quad \text{and} \quad m_k := \inf_{g(k,n)\leq t<g(k+1,n)} X_t$$

can be written as

$$M_k = S_0^{(n)} + Y_k^{(+)} \quad \text{and} \quad m_k = I_0^{(n)} + Y_k^{(-)} \ ,$$

where $\{Y_k^{(+)}\}_{k\geq 0}$ and $\{Y_k^{(-)}\}_{k\geq 0}$ are random walks with the same distribution as $\{X_{g(k,n)}\}_{k\geq 0}$ and independent of $S_0^{(n)}$ and $I_0^{(n)}$ respectively. Since it is clear that

$$m_k \leq X_t \leq M_k \quad \text{for} \quad g(k,n) \leq t < g(k+1,n) \ ,$$

the derivations in Doney (2004) assert that it is possible to 'stochastically' bound the path of X from above and below by two random walks which are equal in distribution to the skeleton proposed in this paper, but with different random starting points.

There is yet another heuristic justification to support the skeleton $\{X_{g(k,n/t)}\}_{k\geq 0}$ as a good random walk approximation of the Lévy process to compute pathwise quantities. The Feynman-Kac representation identifies conditional expectations of functionals of the solution of a Stochastic Differential Equation (SDE) as solutions of a certain Partial Integro Differential Equation (PIDE). We claim that, in some sense, the solution of a Lévy driven SDE sampled over a random grid generated by the arrival times of a Poisson process is equivalent to perform a discretization in time by the method of lines to the associated Feynman-Kac equation. We are not the first to point out this relationship. It was the basis of Carr (1998), where an approximation

for American options of finite maturity is obtained by randomizing the time horizon by an Erlang distribution. Matache *et al.* (2005) also point out informally the relation between a deterministic discretization in time of a Feynman-Kac PIDE and its probabilistic counterpart. Details on this relationship can be found in Ferreiro-Castilla *et al.* (2016).

1.2. Feasibility of the scheme

It is clear from the preceding section that the Euler-Poisson method is of practical interest only if samples from the distribution of Xe(q) are available. In general, there is no reason why the latter distribution is easier to handle than the distribution of X1 itself, needed to set up a plain Euler scheme for a the Lévy process. Nevertheless, recent developments in Wiener-Hopf theory for 1-dimensional Lévy processes have provided a rich enough variety of examples for which the necessary distributional sampling can be performed and thus the Euler-Poisson scheme may lead to simpler numerical techniques. This family of processes are called meromorphic Lévy processes, see Kuznetsov *et al.* (2011) and (2012). For the class of meromorphic Lévy processes, the Wiener-Hopf factors are explicit and hence we can efficiently sample from the distribution of Xe(q) through its factorization. One of the main advantages of the Euler-Poisson algorithm is its robustness with respect to the jump structure, see for example Ferreiro-Castilla *et al.* (2014). Recall that many numerical algorithms to approximate the path of a Lévy process depend on the structure of the jump component of (1).

2. APPLICATIONS

Let us now give a taste of what you can do with the Euler-Poisson scheme. For a comprehensive exposition of the scheme and a range of applications we refer the reader to the papers mentioned in the abstract. Assume then that we have an algorithm to perform the Wiener-Hopf factorization of a Lévy process, then we can keep track of pairs of the type $\{(X_{g(k,n/t)}, \overline{X}_{g(k,n/t)})\}_{k\geq 0}$ or $\{(X_{g(k,n/t)}, \underline{X}_{g(k,n/t)})\}_{k\geq 0}$. It is worth noting that the theory does not allow to keep track of the exact supremum and the infimum of the process simultaneously.

2.1. Computing the first passage time

We will give an example of how to use the scheme to compute the first passage time over a level $u > 0$, i.e.

$$\tau_u := \inf\{t > 0 \,|\, X_t > u\} \,.$$

Assume you can compute the Wiener-Hopf factorization for X and keep track of the running supremum, ie the sequence $\{\overline{X}_{g(k,n/t)}\}_{k\geq 0}$, then

$$k_X^{(n)} := \inf\{k \in \{0,\ldots,n\} \,|\, \overline{X}_{g(k,n/t)} > u\}$$

is an approximation of τ_u. Indeed,

Theorem 2.1. (Ferreiro-Castilla and Van Schaik, 2015). Using the same notation as above, we have

$$\mathbb{E}\left[\left(\frac{t}{n}(k_X^{(n)} \wedge n) - \tau_u \wedge t\right)^2\right] \leq \frac{2t^2}{n} \,.$$

2.2. Approximation of Lévy driven SDEs

Since we have now an skeleton for the path of a Lévy process, it is now natural to investigate its behaviour in approximating Lévy driven SDEs. Let Y be the strong solution of

$$Y_t = y_0 + \int_0^t a(Y_{s-})\mathrm{dX_s} \qquad \mathrm{t} \in [0, \mathrm{T}] \,. \tag{4}$$

The Euler-Poisson scheme is then given by the discrete Markov chain $\widetilde{Y} := \{\widetilde{Y}_{t_i}\}_{i\geq 0}$ defined recursively by

$$\widetilde{Y}_{t_i} := \widetilde{Y}_{t_{i-1}} + a(\widetilde{Y}_{t_{i-1}})\Delta X_{\mathbf{e}_i(n/T)} \qquad \text{for} \quad i \geq 1 \quad \text{and} \quad \widetilde{Y}_0 := y_0 \,, \tag{5}$$

where $\Delta X_{\mathbf{e}_i} := X_{\mathbf{e}_i(n/T)} - X_{\mathbf{e}_{i-1}(n/T)} \overset{d}{=} X_{\mathbf{e}(n/T)}$ and $t_i := \sum_{j=0}^{i} \mathbf{e}_j(n/T)$. We claim that Yetn is an approximation of YT . Indeed,

Theorem 2.2. (Ferreiro-Castilla *et al.*, 2016). Let X have second finite moments, and a in (4) be a Lipschitz function with constant K. Then

$$\mathbb{E}[|Y_T - \widetilde{Y}_{t_n}|^2] \leq \widetilde{K} n^{-1/2}\,, \tag{6}$$

where $\widetilde{K}$ is a constant depending only on K and T.

ACKNOWLEDGMENTS

Part of this research was supported by a Royal Society Newton International Fellowship.

REFERENCES

Broadie, M., Glasserman, P., and Kou, S.G. (1999). "Connecting Discrete and Continuous Path-Dependent Options". *Finance Stoch.*, 3, 55-82.

Carr, P. (1998). "Randomization and the American Put". *Rev. Fin. Studies*, 11 (3), 597-626.

Doney, R.A. (2004). "Stochastic Bounds for Lévy Processes". *Ann. Probab.*, 32 (2), 1545-1552.

Ferreiro-Castilla, A., Kyprianou, A.E., Scheichl, R., and Suryanarayana, G. (2014). "Multilevel Monte Carlo Simulation for Lévy Processes Based on the Wiener-Hopf Factorization". *Stochastic Process. Appl.*, 124 (2), 985-1010.

Ferreiro-Castilla, A., Kyprianou, A.E., and Scheichl, R. (2016). "An Euler-Poisson Scheme for Numerical Approximation of Lévy Driven SDEs". *J. Appl. Probab.*, 53 (1), to appear.

Ferreiro-Castilla, A., and Van Schaik, K. (2015). "Applying the Wiener-Hopf Monte Carlo Simulation Technique for Lévy Processes to Path Functionals". *J. Appl. Probab.*, 52 (1), to appear.

Kuznetsov, A., Kyprianou, A.E., and Pardo, J.C. (2012). "Meromorphic Lévy Processes and their Fluctuation Identities". *Ann. Appl. Probab.*, 22 (3), 1101-1135.

Kuznetsov, A., Kyprianou, A.E., Pardo, J.C., and Van Schaik, K. (2011). "A Wiener-Hopf Monte Carlo Simulation Technique for Lévy Process". *Ann. Appl. Probab.*, 21 (6), 2171-2190.

Matache, A.M., Nitsche, P.A., and Schwab, C. (2005). "Wavelet Galerkin Pricing of American Options on Lévy Driven Assets". *Quantitative Finance*, 5 (4), 403-424.

35. ON SOME GENERALIZATIONS OF UNIVARIATE NORMAL DISTRIBUTION

J. Filus
Oakton Community College, Department of Mathematics and Computer Science, USA,
jfilus@oakton.edu

ABSTRACT

We consider a class of transformations which produce generalizations of the univariate normal density. Among them are power transformations similar, but not identical, to the Box-Cox transform. We investigate moments and analytical properties of the, so obtained, generalizations of normal density.

1. GENERAL DEFINITIONS

Consider any transformation $\varphi: R^1 \rightarrow R^1$ defined on the whole set of real numbers R1 whose range is also the whole set R^1. Assume also that φ is strictly monotonic with both φ and φ^{-1} having continuous derivative. Moreover we will assume that $\varphi(0) = 0$.

For any such diffeomorphism $R^1 \rightarrow R^1$ we define the following probability density on the set of all real numbers:

$$f(x) = (|\varphi'(x)| / \sigma\sqrt{2\pi}) \exp[- |\varphi(x)|^2 / 2\sigma^2] \quad (1)$$

where $|\varphi'(x)|$ is absolute value of derivative of the $\varphi(x)$.

Realize that the identity '$\varphi(x) = x$' satisfies all the above properties so it belongs to the defined class of the functions φ. Substituting the identity x into (1) one obtains as special case of (1) the familiar normal $N(0, \sigma)$ density

$$f_0(x) = (1 /\sigma\sqrt{2\pi}) \exp[- x^2 / 2\sigma^2] \quad (2)$$

Thus (1) is considered as generalization of the N(0, σ) density (2).

Realize that the density (1) is symmetric whenever the defining diffeomorphism satisfies $\varphi(- x) = - \varphi(x)$, and $|\varphi'(-x)| = |\varphi'(x)|$, for any x. For example, this is the case for the cubic function

$\varphi(x) = Ax^3$ with a real parameter $A \neq 0$.

Notice that to extend the class of the densities (1) to be the extension of the normal N(μ, σ) rather than N(0, σ) we may consider the functions

$\varphi(x- \mu)$ instead of $\varphi(x)$ for an arbitrary real μ.

This difference, however, is not essential for the present considerations.

Remark. The assumptions defining class of the functions φ may be relaxed. As for its domain the whole line R^1 may be replaced by any bounded open interval, say $(a, b) \subseteq R^1$ while the range would remain the whole R^1.

As an example, consider $\varphi(x) = \tan(x)$ as defined on the open interval $(-\pi/2, \pi/2)$.

The following probability density is only defined on this interval:

$$f(x) = (\sec^2(x) /\sigma\sqrt{2\pi}) \exp[- \tan^2(x) / 2\sigma^2] \quad (3)$$

The above can also be considered as generalization of the normal N(0, σ), this time on the bounded interval $(-\pi/2, \pi/2)$. In other words, random variable $U = \tan(X)$ has the normal distribution, while the so defined random variable X has density on $(-\pi/2, \pi/2)$, given by (3). This "tangent-normal" density is not the only example of such densities defined on open bounded intervals (other easy example would be "cotangent-normal" density defined on $(0, \pi)$).

2. POWER CASE

A. Definitions

As special case of the transformations φ consider the power transformations defined as follows:

$u = x^{\alpha}$ for $x \geq 0$ and

$u = -(-x)^{\alpha}$ for $x < 0$, (4)

where $-\infty < x < \infty$ and $\alpha > 0$ and therefore $-\infty < u < \infty$.

Suppose X, U are random variables related by (4).

If U has the normal $N(0, \sigma)$ density then we declare X to be "normal-Weibull $NW(0, \sigma; \alpha)$" distribution whose density has the following form:

$$f(x; 0, \sigma, \alpha) = (\alpha|x|^{\alpha -1} /\sigma\sqrt{2\pi}) \exp [- |x|^{2\alpha} / 2\sigma^2] \quad (5)$$

where |x| denotes the absolute value of x.

The density (5) is derived from the normal in a similar way as the Weibull density was derived from the exponential. This resemblance justifies the name 'normal-Weibull' for the density (5) which is slightly *different* than the power-normal distribution which follows **Box-Cox** power transformation (Box and Cox, 1964) defined for $x \geq 0$ only (but for all real α so it is not always a diffeomorphism).

Realize that upon the substitutions (4) the integral

$$\int_{-\infty}^{\infty}(\alpha|x|^{\alpha -1} /\sigma\sqrt{2\pi}) \exp [- |x|^{2\alpha} / 2\sigma^2]dx \text{ reduces to the integral}$$
$$\int_{-\infty}^{\infty}(1/\sigma\sqrt{2\pi})\exp[- u^2 / 2\sigma^2] du \quad (6)$$

and so equals to one.

As in general case, considered in Section 2, somewhat more general than (4) is the following transformation:

$u = |(x-\mu)|^{\alpha}$ for $x \geq \mu$ and

$u = - |(x-\mu)|^{\alpha}$ for $x < \mu$, (7)

From the latter one can obtain slightly more general density than (5) assuming that the random variables X, U are related by (7) and that U has normal $N(0, \sigma)$ distribution.

This density has the form:

$$f(x; \mu, \sigma, \alpha) = (\alpha|x - \mu|^{\alpha -1} /\sigma\sqrt{2\pi}) \exp [- |x - \mu|^{2\alpha} / 2\sigma^2] \quad (8)$$

Obviously, with $\mu = 0$ one obtains back (4) and the density (5).

A slightly different case one obtains by setting

$U - \mu = |X|^{\alpha}$ for $U \geq \mu$ and $X \geq 0$,

$U - \mu = - |X|$ for $U \leftarrow \mu$ and $X < 0$ (9)

with U having $N(\mu, \sigma)$ distribution.

B. Modified Power Case

Looking at the form of the density (5) one realizes that if $\alpha > 1$ then the density takes on the value zero at $x = 0$, so the minimum, while the corresponding normal density $N(0; \sigma)$ takes on the maximum. This fact is rather not desirable so one may try to find a way out from this situation by replacing the power transformation (4) by the following one:

$u = x^{\alpha} + x$ for $x \geq 0$, and

$u = - |x|^{\alpha} + x$ for $x < 0$, (10)

Realize that, the so defined, transformation (10) satisfies all the conditions for the diffeomorphisms φ: $R^1 \rightarrow R^1$ as specified in Section 2.

Moreover the φ given by (10) satisfies $\varphi(-x) = -\varphi(x)$, for all x.

C. The Moments

The expected value of the density (5) is zero as its graph is symmetric about the y-axis i.e., $f(-x; 0, \sigma, \alpha) = f(x; 0, \sigma, \alpha)$. Generally, all the moments for (5) can be obtained in terms of the Mellin function for $N(0, \sigma)$ density, if σ is known. Namely, the integral

$$\int_{-\infty}^{\infty} x^k \, (\alpha|x|^{\alpha-1} / \sigma\sqrt{2\pi}) \exp[-|x|^{2\alpha} / 2\sigma^2]dx$$

can be reduced to the following sum of integrals

$$\int_0^{\infty} u^{k/\alpha} \, (1/\sigma\sqrt{2\pi})\exp[-u^2/2\sigma^2]\, du + \int_{-\infty}^{0} (-1)^k(-u)^{k/\alpha} \, (1/\sigma\sqrt{2\pi})\exp[-u^2 / 2\sigma^2]du.$$

The latter sum equals zero for each k = 1, 3, ... and for k = 2, 4, ... equals

$M(k/\alpha + 1)$, where M(z) is the Mellin function from the normal $N(0, \sigma)$ distribution.

For example, the second moment i.e., the variance of X equals $M(2/\alpha + 1)$. As for the absolute moments of X they simply are given by

$$\int_{-\infty}^{\infty} |x|^k \, (\alpha|x|^{\alpha-1} / \sigma\sqrt{2\pi}) \exp[-|x|^{2\alpha} / 2\sigma^2]dx$$

for each real positive k and equal to the k/α-th absolute moment of $N(0, \sigma)$ distribution.

If $\alpha = 1$ then density (5) reduces to the normal $N(0, \sigma)$ density. Other special cases of an interest are $\alpha = 2$ and $\alpha = ½$.

In the first case (5) takes its particular form:

$$f(x; 0, \sigma, 2) = (2|x| / \sigma\sqrt{2\pi}) \exp[-|x|^4 / 2\sigma^2] \qquad (11)$$

The k-th moment of latter density is M(k/2 + 1), where M(z) is the Mellin function for N(0, σ) density. In particular, since the expectation of (5) is zero, the variance equals M(2).

In the second case, one obtains the following version of (5):

$f(x; 0, \sigma, ½) = (1 / 2\sqrt{|x|}\sigma\sqrt{2\pi}) \exp [- |x| / 2\sigma^2]$.

Here, consequently, the k-th moment is given as M(2k +1) and so the first two moments are M(3) and M(5), respectively.

3. ON STATISTICAL INFERENCE

Statistical analysis for the density (1) is significantly facilitated when the defining diffeomorphism φ is known, given, for example, by formula (4). Thus, in such cases, any sample $x_1, x_2, \ldots x_n$ from the random variable X can be transformed into the sample $u_1, u_2, \ldots u_n$ from the normal variable U, where for i = 1, 2, ... , n $u_i = \varphi(x_i)$. The fit of the data $x_1, x_2, \ldots x_n$ to the model (1) is then determined by fit of the sample $u_1, u_2, \ldots u_n$ to the usual normal density. In case of the transformation (4), only a knowledge of the parameter α is needed to transport the sample from X distributed according to (5) into the sample from the normal pdf.

Conversely, to sample data from density (1), one only needs to generate data from the normal pdf of U, whenever the transformation φ is at a disposal.

REFERENCES

Box, G.E.P., and Cox, D.R. (1964). "An Analysis of Transformations". *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

36. CONSTRUCTION OF CONDITIONAL PROBABILITY DENSITIES AS ENFORCED REGRESSION

J. Filus
Oakton Community College, Department of Mathematics and Computer Science, USA,
jfilus@oakton.edu

L. Filus
Northeastern Illinois University, Department of Mathematics, USA,
L-Filus@neiu.edu

ABSTRACT

Let Y be a random variable of the main interest and random variables X_1, ... ,X_k serve as its explanatory variables. We use our earlier "method of parameter dependence" applied to construction of various multivariate probability distributions. This time we restrict the method to construction of the conditional distribution of the Y, given realizations x_1, ... ,x_k of the explanatory variables X_1, ... , X_k. Relative easiness of the latter constructions prompted us to apply the obtained large class of the conditional distributions for the existing regression theory as alternative models. These alternative models are intended to extend those well known usually given as the conditional expectation of, say, Y, given x_1, ... ,x_k. Apparently, the conditional distributions bring much more statistical information than the corresponding expectations alone. Most of these models are not yet known in literature. We will present them, discuss and analyze their analytic properties.

1. INTRODUCTION

We apply the method of parameter dependence (Filus & Filus, 2012, 2013) used for the construction of a variety of multivariate probability distributions, to analyze closer the ("weak") dependence of a given random variable Y from a set of explanatory random variables, say, $X_1, \ldots, X_k$. High generality of the method of parameter dependence opens the possibility to present it now as a base for settings of the theory that, in a sense, extends and modifies the theory of classical regression and the underlying models. Recall that, perhaps up to the presence, the vast majority of (stochastic) models and methods, employed in the classical regression theory and its applications, were essentially related to the conditional expectation

$$E[\, Y \mid x_1, \ldots, x_k \,] \tag{1}$$

of a random variable (or a random vector) of the main interest Y, given realizations $x_1, \ldots, x_k$ of a set of explanatory random variables $X_1, \ldots, X_k$. This kind of stochastic model for the statistical theory of regression is limited with respect to an amount of statistical information and thus the precision of the corresponding predictions is limited as well.

Obviously, a better (i.e., more precise) model compared to (1) would be the stochastic model in the form of a whole conditional probability distribution (not just its expected value) of the same random variable Y, given the same occurrences of the explanatory values $x_1, \ldots, x_k$.

A practitioner statistician would certainly be "more happy" having the possibility to enrich the typical regression model (1) by the following conditional cdf:

$$P(Y \leq y) = G_k(y \mid x_1, \ldots, x_k) \tag{2}$$

or, whenever it exists, the corresponding conditional pdf:

$$g_k(y \mid x_1, \ldots, x_k). \tag{2*}$$

As model (1) is "part" of the (2) or (2*) model, (2) as well as its theory, is the, proposed here, natural extension of this part of the classical regression theory which deals with (1) - like models. Of course, this possibility was a very well-known fact since a long time. But using the stochastic dependence models in the form (2) and (2*) was (possibly, only, with exception of the conditionals associated with the classical multivariate normal pdfs) seen as "not realistic". The reason for this shortcoming is explained as follows. In almost all non-Gaussian situations there were no well defined and handy enough analytical forms of the conditional distributions (2) that would explicitly and meaningfully reflect the stochastic dependences of the random variables Y from the explanatory ones X_1, ... ,X_k. Discovering the pattern (of the parameter dependence) (Filus and Filus, 2012, 2013) for an explicit description of the stochastic dependences and, associated with it, new possibilities for construction of the conditional distributions (2) on a "mass-scale", gave us hope to enrich the statistical regression theory in a new way.

The main goal for all that is an expected (in many practical cases) improvement in accuracy of predictions for the phenomena so far modeled by regular (conditional expectation) regression methods. At this point let us recall the known fact that some (to a measure similar) results were obtained and the corresponding theory exists in the literature. This theory, known as the 'quantile regression methods' was first introduced by Koenker and Bassett (1978), and is continued by some authors up to now. Similar to ours, this approach is aimed to complement the classical linear regression analysis, mainly by replacing the typical investigations centered on the conditional expectation E[Y $|x_1$, ... ,x_k] by more general considerations on the conditional quantiles directly associated with the (whole) conditional probability distribution. The main advantage of the extension is the possibility to obtain more statistical information than by using traditional methods of regression. The essential difference between those methods and our present approach mainly relies on the parametric character of the presented extended regression-models. In practice this means that our models are capable of "embracing" (in one analytic formula) the probabilities P ($Y \leq y$ $|x_1$, ... ,x_k)) for 'all the y' values "at the same time", not just for 'each' quantile y separately, which, practically, reduces the investigations to few values of y, only.

As for the multivariate normal case exception, it should be admitted that in the extended regression theory presented here, both the common restrictions on the

normality and the linearity of the regression function, can be relaxed in a natural way.

The general pattern for our stochastic dependence description is obtained by means of some method of conditioning. This method relies on a kind of randomization of the originally constant parameter(s) (the baseline, i.e., "no explanatory random variable's influence case") of any pdf $g_0(y)$ of the random variable Y. Namely, we consider these parameters as continuous (and in general nonlinear) function of the independent explanatory random variables $X_1, \ldots, X_k$, each of them having some known probability distribution. For example, Y may be considered as the life time of an object and $X_1, \ldots, X_k$ may be extra stresses put on this object.

This procedure yields the determination of a set of classes of conditional pdfs $g_k(y|x_1, \ldots, x_k)$ or cdfs $G_k(y|x_1, \ldots, x_k)$ of the random variable Y, given the realizations of the independent (explanatory) random variables $X_1, \ldots, X_k$. Our purpose is to provide a method for explicit determination of any conditional probability $P(a \leq Y \leq b \mid x_1, \ldots, x_k)$ for all: $-\infty \leq a < b \leq \infty$, in a concise analytical form as a function of the realizations $x_1, \ldots, x_k$ of the explanatory random variables $X_1, \ldots, X_k$. In particular, we consider conditional survival functions $P(Y \geq y|x_1, \ldots, x_k)$ with Y interpreted as the life time of an object and $X_1, \ldots, X_k$ as stresses that the object endures.

The above probabilities that the object would survive at least time period of length y (for any y) is an analytically given continuous function, say $r(x_1, \ldots, x_k)$ of the stresses' realizations $x_1, \ldots, x_k$. This function we call the "enforced regression function of the second kind", and this is the main stochastic model we construct in this paper. We precede this by the notion of "enforced regression of the first kind" which also covers conditional expectations $E[Y|x_1, \ldots, x_k]$ when the baseline probability distribution of Y is neither normal nor exponential. When Y is normally distributed, the enforced regression of the first kind becomes identical with the ordinary (not necessarily linear) regression function. Both first and second kind of the enforced regression are based on the parameter dependence (also called 'weak dependence') between the random variable Y and the set of random variables $X_1, \ldots, X_k$. Between others, we explain this kind of stochastic dependence in the wider framework of joint multivariate distributions.

2. DESCRIPTION OF THE PROBLEM

We are interested in the probability distribution of a random variable Y which, in particular, may correspond to the (random) life-time of an object. In case of actuary applications, Y may be considered as the residual life- time of a client whose age at the moment of registration is t . Assume that in the "regular situation" the residual life-time Y has a given known probability density function $f(y ; \theta)$, where θ is its scalar or vector parameter. In some significant cases the situation complicates because of the presence of extra "disturbing facts" taking place when potential clients were subjected to special stresses that could affect their residual life-time. In some cases the amount of a given stress can be measured by a (possibly random) quantity X_i. For example, X_i can be the time the stress was endured, multiplied by its intensity, for example as the amount of nicotine or alcohol consumed in an average day. Thus, the "disturbing factors" having an impact on the residual life-time Y (probability distribution) may be considered as either a single random variable X or as a set of such (independent) random variables $X_1, X_2, \ldots , X_k$, $k = 2, 3, \ldots$.

We aim to find an efficient way to describe, analytically, the general stochastic mechanism by which a random stress X "influences" the random life-time Y. In the general version (of the "extended regression theory") the question can be formulated as: "how can one analytically express the influence of "any" random variable X on another random variable Y' probability distribution, when the two are not independent and with no given explicit algebraic relation (transformation) between them".

The stochastic (indirect) influence of a value x, corresponding to the random event (observation) $X = x$, on possible values y of the random variables Y (the stochastic impact of x on y) is understood as the impact of x on the probability (or probability density or hazard rate) of occurring y , when $X = x$ happens. This, in general, means that different values x, x* of X may bring different probabilities (or densities) of the given random event $Y = y$. Speaking more generally, in the considered setting, the random event $X = x$, "influences" the probabilities $P (Y \leq y)$, for any fixed real value y.

In other words, we are seeking for the conditional probability distributions $P (Y \leq y \mid X = x)$, or the corresponding conditional probability densities, hazard rates, etc.

In the framework we have chosen, the phrase: "Size of an 'incentive' (or, in particular, 'stress') x changes the probability density f (y ;θ) of the life-time Y or (equivalently), the x " changes Y's hazard (failure) rate λ(y ; θ) through a change of the parameter θ" will be understood as: "The stress X changes, (proportionally to its magnitude x) the numerical value of the parameter(s) θ". In other words, "a change in θ is 'proportional' to a magnitude x of the stress X" , given the random event X = x. To describe the (deterministic) relationship between θ and x one can impose, as new values of θs , a 'hypothetical' continuous (not necessarily linear) function of x : θ = θ(x) . For example, we may let $\theta \rightarrow \theta(x) = \theta$ 'times' $(1 + ax + bx^2)$ and then, assuming such a model, statistically estimate the (new) parameters a and b , where the factor θ on the right hand side of the above equality may be known from "previous" (baseline) estimation procedures. We obtain the conditional density $g_2(y \mid x)$ of Y , given X = x , as defined by $g_2(y \mid x) = f(y ; \theta(x))$.

The last key formula yields directly the extended regression theory formulation.

Equivalently, one can define the conditional hazard (failure) rate by transforming the original hazard rate $\lambda_2(y , \theta)$ of Y to the conditional one

$\lambda_2(y \mid x) = \lambda_2(y , \theta(x))$.

3. CONCLUSIONS

Having defined the conditional pdf $g_2(y \mid x)$ and the marginal pdf $g_1(x)$ of the stress X, we automatically obtain the joint pdf h(x, y) of the random variables X and Y simply as the arithmetic product

$h(x, y) = g_2(y \mid x)\, g_1(x)$

(when the random vector (X, Y) is meaningful in the investigations).

Anyway this, apparently, is the method (also called the 'parameter dependence') for constructing bivariate and multivariate probability densities. Being a general method of construction it can be considered as an extension of the paradigm on which the

construction of classical bivariate and multivariate Gaussian densities is based, see Filus and Filus (2013). The same applies to the associated Gaussian conditional densities. So, the method of parameter dependence can be viewed as an extension of the method of how the classical multivariate normal were constructed. The same statement may apply to the, here considered, enforced regression. We will explain this closer in the presentation.

REFERENCES

Filus, J.K., and Filus, L. Z. (2013). "A Method for Multivariate Probability Distributions Construction via Parameter Dependence". *Communications in Statistics: Theory and Methods*, Vol. 42, no. 4, 15, 716-721.

Filus, J.K., and Filus, L. Z. (2012). "Multivariate 'Pseudodistributions' as Natural Extension of the Multivariate Normal Density Pattern – Theory". *American Institute of Physics Conference Proceedings 1479, Numerical Analysis and Applied Mathematics*, vol. 1479, 1417-1420.

Koenker, R., and Bassett, G.W. (1978). "Regression Quantiles". *Econometrica*, 46, 33-50.

37. RISK OF DATA LOSS ON QR CODES: HADAMARD MATRICES AND LINKS TO BLOCK DESIGNS

C. Francisco
Universidade Aberta, Departamento de Ciências e Tecnologia, Portugal,
carlasusete@hotmail.com

T.A. Oliveira
Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal,
teresa.oliveira@uab.pt

ABSTRACT

As a branch of Block Designs, research on Balanced Incomplete Block Designs (BIBD) arose several interesting and defying problems within Combinatory Mathematics.

Hadamard Matrices are present in our daily life, they give rise to a class of block designs named Hadamard configurations. It is easy and current to find their different applications based on new technologies and codes of figures such as Quick Response Codes (QR Codes). These are bi-dimensional barcodes that can be easily read by common devices which have image capture function, such as mobile phones.

Risk is the potential of losing something of value, weighed against the potential to gain something of value. There are several types of risk, one of them is the information loss for QR Codes and that will be presented in this paper. Connections between the various methodologies and QR Codes will be discussed.

1. INTRODUCTION

Research on Block Designs began with randomized Block Designs which play an important role in the design and analysis of experiments, see e.g., Calinski and Kageyama (2000). A very important class of Block Designs is known as Balanced Incomplete Block Designs. The adjective incomplete refers to the block size (k) and the variety set (v), with $k < v$, while balanced refers to each pair of varieties occurring in the same number of blocks, Oliveira *et al.* (2006). In this designs every pair of varieties concur in the same number of blocks. There is a strong connection between BIBD and Hadamard Matrices.

Hadamard Matrices are a class of matrices introduced by Jacques Hadamard in 1893, see Hadamard, J. (1893). In this class of matrices, entries are $+\mathbf{1}$ and $-\mathbf{1}$ so that rows are mutually orthogonal. A Hadamard Matrix, **H,** is said to be skewed if the relation $\mathbf{H}^{\mathbf{T}} + \mathbf{H} = \mathbf{2I}$ holds, where $\mathbf{H}^{\mathbf{T}}$ stands for the matrix transpose of **H** and **I** represents the identity matrix. Besides this condition, the matrix rows are mutually orthogonal, which will have direct impact to their use, namely in geometric terms, this means that every pair of rows, $(\boldsymbol{i}, \boldsymbol{j})$, with $\boldsymbol{i} \neq \boldsymbol{j}$, will represent two perpendicular vectors.

QR Codes are two-dimensional barcodes that can be easily read by a device capable of image capture, as is the case of most existing mobile phones. These codes can represent text, an address for a web site (URL), a phone number, a geo-referenced location, an email, a contact or an SMS. They were initially used for cataloguing parts in vehicle production. Nowadays many QR codes are used for several different tasks such as inventory management and stock control, from industry to trade. QR Codes are common in magazines and advertisements, where they are used to encode different kinds of information, or for instance to provide contact information, as detailed as if they were a personal or business card. QR Codes have become a focus of advertising companies, since they provide an effortless way to access the advertised brand's website. QR codes have, as the base of their error correcting structure, Hadamard Matrices.

2. OBJECTIVES

This paper proposes to illustrate error correction on QR Codes through the use of Hadamard Matrices as error correction promoters and their connection to incomplete Block Designs. Hadamard Matrices form an important part of the family of error correction codes, denominated ReedMuller codes used in practice on the QR Codes, CDs, DVDs, DSL Internet, digital tape devices, digital TV and satellite transmission of signals. This paper will focus in particular on the error correction applied to QR Codes.

3. HADAMARD MATRICES

A Hadamard Matrix is a square matrix $\mathbf{H}_n$ of order n with entries $\pm$ if $\mathbf{H}_n \mathbf{H}_n^{\mathrm{T}} = n\mathbf{I}_n$. If $\mathbf{H}_n$ is Hadamard Matrix then $\mathbf{H}_n \mathbf{H}_n^{\mathrm{T}} = n\mathbf{I}_n$. A Hadamard Matrix remains so when any row or column is multiplied by $-\mathbf{1}$. If $\mathbf{H}_n$ exists for n = 1 then $\mathbf{H}_2$, can be written: $\mathbf{H}_2 = \begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & -\mathbf{1} \end{bmatrix}$.

The necessary condition for the existence of a Hadamard Matrix $\mathbf{H}_n$, $n > 2$ is that $n \equiv 0 \ (mod\ 4)$; more about this can be found on, Hall (1986).

Whether this necessary condition is sufficient as well, it is not known yet.

4. RELATIONSHIP BETWEEN HADAMARD MATRICES AND BIBD

Consider a Hadamard Matrix $\mathbf{H}_{4u}$, which without loss of generality, is assumed to be in its normal form. Delete from $\mathbf{H}_{4u}$, its first row and first column of all ones to obtain a matrix **A** of order $(4_u - 1) \times (4_u - 1)$.

Define, $\mathbf{N} = \frac{1}{2}(\mathbf{A} + \mathbf{J}_{4u-1})$.

This means that **M** is obtained from **A**, by replacing the –1's in A by zero and keeping +1's unaltered. Then, it is not hard to see that **N** is the incidence matrix of a BIB Design with parameters:

$$v = 4u - 1 = b;\ r = 2u - 1 = k;\ \lambda = u - 1 \tag{1}$$

Conversely, if **M** is the incidence matrix of a BIB Design with parameters given by (1), then by replacing the zeros in **M** by **–1** and bordering the resultant matrix by a row and column of all ones, one gets a Hadamard Matrix of order ***4u***. We thus have the following Theorem:

Theorem 1. The existence of a Hadamard Matrix of order ***4u*** is equivalent to the existence of a BIB Design with parameters given by (1).

Example 1. Consider a Hadamard Matrix $\mathbf{H}_{16}$ which can be obtained by forming the tensor product $\mathbf{H}_4 \otimes \mathbf{H}_4$, where $\mathbf{H}_4$ is as bellow:

$$\mathbf{H_4} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{1} \end{bmatrix}$$

Following the construction method described above, we get a solution of a BIB Design with parameters $v = 15 = b,\ r = 7 = k,\ \lambda = 3$.

5. APPLICATION OF HADAMARD MATRICES TO ERROR CORRECTION

Reed-Muller codes are a family of linear error-correcting codes used in communications. Special cases of Reed-Muller codes include the Hadamard code, the Walsh-Hadamard code, and the Reed-Solomon code.

The Hadamard code is an error-correcting code that is used for error detection and correction when transmitting messages over very noisy or unreliable channels. A famous application of the Hadamard code was the NASA space probe Mariner 9 in 1971, where the code was used to transmit photos of Mars back to Earth. Generalized Hadamard codes are obtained from an $\boldsymbol{n} \times \boldsymbol{n}$ Hadamard matrix **H**. In particular, the $\mathbf{2}\boldsymbol{n}$ codewords of the code are the rows of **H** and the rows of **–H**.

The higher the error correction level, the less the available storage capacity there is. Reed-Solomon codes are of the same family of error correcting codes as the

Hadamard codes. Reed-Solomon codes are closely related to Hadamard Matrices as well. In recent papers, see Francisco and Oliveira (2014), the author's paid particular attention to the importance of links to Block Designs and their application on Risk of Data Loss.

6. RISK

Risk is a probability or threat of damage, injury, liability, loss, or any other negative occurrence that is caused by external or internal vulnerabilities, and that might be avoided through a pre-emptive action. There are several types of risk, meaning different classes or various forms of risk such as, project risk which are the factors that could cause a project to fail, business risk which are associated with the level of exposure a business will face if a project fails, production system risk which considers the running costs of a project, economic risk which can be manifested in lower incomes or higher expenditures than expected, health, safety, and environmental risk. Other types of risk are security risk which concerns the protection of assets from harm caused by deliberate acts and information technology, information security risk, among many others. The one that concerns QR Codes is the information security risk in its area of information loss risk. Here, information loss risk refers to the problematic of losing important data, contained in the QR Code.

Risk assessment is the process of identifying potential hazards and analyse what could happen if a hazard occurs. A hazard is a recognized threat. Therefore risk assessment is the determination of quantitative or qualitative value of risk related to a concrete situation and a recognized threat. There are various methods for performing risk assessment, however there is one believed to be more recognized for most organisations. This method consists of five steps: identify the hazards, decide who might be harmed and how, evaluate the risks and decide on precaution, record the findings and implement them and the final step is to review the assessment and update it if necessary. Quantitative risk assessments include a calculation of the single loss expectancy (SLE), of a resource. Single loss expectancy is defined as the monetary value expected from the occurrence of a risk on an asset and is mathematically defined as:

SLE = Asset Value x Exposure Factor

The exposure factor is represented in the impact of the risk over the considered asset, or as a percentage of asset lost. Illustrating with an example, if the asset value is reduced by one third, the exposure factor value is .33. If the asset is completely lost, the Exposure Factor is 1.0 (meaning 100%).The result of the previous formula is a monetary value, represented in the same unit as the Single Loss Expectancy is expressed, euros, dollars, yens, etc. Information assurance is the practice of assuring information and managing risks related to the use, processing, storage, and transmission of information or data and the systems and processes used for those purposes. For QR Codes there is a significant risk of damage and therefore risk of loss of the information contained in it. To prevent the loss of information on QR Codes a greater error correction level should be used when the code is created.

There are four levels of error correction that can be used: Level L up to 7% damage, Level M up to 15% damage, Level Q up to 25% damage and Level H up to 30% damage. In this conditions where the QR Code is exposed to the elements, such as rain, wind, sun, dirt, etc. the H level should be used, so even though there was 30% of damage to this QR Code, the information contained in it would still be totally readable. As such, when creating QR Codes for example for business cards, the chances of it getting dirty are rather low and it is desirable to have a QR code as small as possible. Here a standard 7% error threshold is enough and very suitable for cards and other printed matter. However if QR codes are going to be used outdoors, on cars and trucks or any place where data may be obscured by dirt then setting the error correction to a higher setting is advisable. Unfortunately, the higher the error correction level that is used, the less information that can be written on to the QR Code. So, there is a compromise between risk mitigation and information providing.

The QR codes in figure 1, contain the url to the Icra 6 & Risk 2015 Conference. The one, at the right is damaged but still is readable thanks to the Hadamard Matrices:

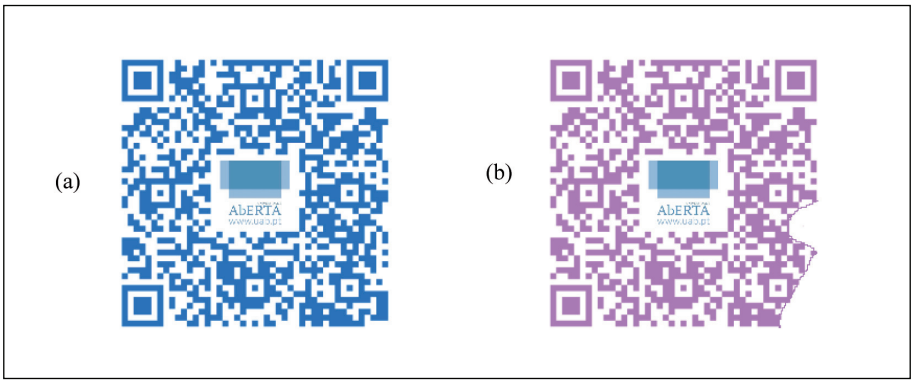


Figure 1. (a) QR code to http://www.uoc.edu/portal/en/symposia/icra6/Home/; (b) damaged QR code to http://www.uoc.edu/portal/en/symposia/icra6/Home/

7. FINAL CONSIDERATIONS

Hadamard Properties can minimize the risk of Data Loss based on the information contained in the damaged area. BIBD can be obtained throughout Hadamard Matrices and many other results explored in the Experimental Design can be of great utility in this field.

ACKNOWLEDGEMENTS

This research was partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the project PEst-OE/MAT/UI0006/2014.

REFERENCES

Calinski, T., and Kageyama, S. (2000). *Block Designs: A Randomization Approach*. Vol. I: Analysis. Springer, 150.

Francisco, C., and Oliveira, T. A. (2014). *BIBD, Hadamard Matrices and Risk of Data Loss on QR Codes, in Proceedings of SMTDA 2014, 3rd Stochastic Modeling*

Techniques and Data Analysis. International Conference in Lisbon, Portugal, 11-14 June 2014. Book of Abstracts, 61.

Hadamard, J. (1893). "Résolution d'une question relative aux determinants". *Bulletin des Sciences Mathématiques* 17, 240–246.

Hall Jr, M. (1986). *Combinatorial Theory*, 2nd edition. New York, Wiley.

Oliveira, T.A., Ceranka, B, and Graczyk, M. (2006). "The Variance of the Difference of Block Effects in the Balanced Incomplete Block Designs with Repeated Blocks", *Colloquium Biometryczne*, 36, 115-124.

38. COMPONENTS WITH HIGHER AND LOWER RISK IN A RELIABILITY SYSTEM

J. Freixas and M. Pons
Universitat Politècnica de Catalunya (Campus Manresa),
Departament de Matemàtica Aplicada 3, Spain,
josep.freixas@upc.edu;
montserrat.pons@upc.edu.

ABSTRACT

A new reliability importance measure for components in a system, that we call Representativeness measure, is introduced. It evaluates to which extent the performance of a component is representative of the performance of the whole system. Its relationship with Birnbaum's measure is analyzed, and the ranking of components given by both measures are compared. These rankings happen to be equal when all components have the same reliability but different in general. In contrast with Birnbaum's, the Representativeness reliability importance measure of a component does depend on its reliability.

1. INTRODUCTION

Many reliability importance measures (RIM) have been defined up to now, see Kuo and Zhu (2012) for a complete and detailed survey on importance measures. The most well–known, proposed by Birnbaum (1969), evaluates the contribution of a component to the system reliability by the rate at which the system reliability improves as the reliability of the component improves. In this work we propose a new index for measuring the reliability importance of a component. This index evaluates the degree of "coupling" between one component and the whole system, i.e., the likelihood that the system and the component are in the same state. We call it Representativeness Measure

because by selecting and analyzing the performance of components with a high value of this measure we can have an acceptable estimation of the performance of the system. In this sense we think that this measure can be useful for maintenance purposes. This measure gives rise to an associated structural importance measure when all components have the same reliability equal to one half.

In section 2 the context of the work is established and some basic definitions are given. The Representativeness RIM and SIM are defined in section 3 and their main properties are proved. In section 4 we compare these new measures with Birnbaum measures. All proofs of the different statements are omitted but available on request.

Notations:

N set of components, which are assumed to be numbered consecutively from 1 to n.

x_i binary random variable associated to component i: $x_i = 1$ if i is functioning, $x_i = 0$ otherwise.

$\phi(\mathrm{x})$ binary structure function of the system: $\phi(\mathrm{x})$ = 1 if system is functioning, $\phi(\mathrm{x})$ = 0 otherwise.

p_i reliability of the ith component, $Pr\{x_i = 1\}$.

$\mathbf{p}$ reliability vector, $\mathbf{p} = (\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_n)$.

$(a_i, \mathbf{p})$ vector $\mathbf{p}$ with the ith component of p changed to a.

π collection of path sets of the system.

(N, π) system defined by the collection π of path sets.

$h(\mathbf{p})$ system reliability, $Pr\{\phi(\mathrm{x}) = 1\}$.

$I_i^B(\mathbf{p})$ Birnbaum RIM of component i.

$I_i^{Bs}(\mathbf{p})$ Birnbaum SIM of component i.

$I_i^{R}(\mathbf{p})$ Representativeness RIM of component i.

$I_i^{Rs}(\mathbf{p})$ Representativeness SIM of component i.

Assumptions:

1. the random variables x_i are mutually statistically independent.

2. the systems we consider are coherent:

    - the structure function ϕ is nondecreasing,

    - all components are relevant.

3. $0 < p_i < 1$ for any component i.

2. FUNDAMENTALS

We consider binary systems, i.e., systems in which there is a random variable associated to each component which takes the value 1 if it is functioning and 0 otherwise. These random variables are assumed to be mutually statistically independent. The structure function of the system is a binary function of these random variables that takes the value 1 if the system is functioning and 0 otherwise. We assume that this structure function is nondecreasing. The structure function of a system can be expressed in terms of its path sets.

This is why we usually denote a system by (N, π), where $N = \{1, 2, \ldots, n\}$ is the set of components and π is the collection of path sets. Recall that a path set is a set of components which by functioning ensure that the system is functioning. A path set is minimal if it does not have proper path subsets. A component i is relevant if it belongs to at least one minimal path set.

The systems we consider have a nondecreasing structure function and all the components are relevant. They are referred to as coherent systems. The reliability $h(\mathbf{p})$ of the system, i.e., the probability of it being functioning, depends on its structure function and on the reliability of each component. If p_i is the reliability of component i (the probability of it being functioning), and p is the reliability vector $(p_1, p_2, \ldots, p_n)$, then the system reliability can be expressed by means of path sets:

$$h(\mathbf{p}) = \sum_{S \in \pi} \prod_{i \in S} p_i \prod_{i \in N \setminus S} (1 - p_i).$$

Recall also that, by a conditional probability argument, for any component i we get:

$$h(\mathbf{p}) = p_i h(1_i, \mathbf{p}) + (1 - p_i) h(0_i, \mathbf{p})$$

and from this expression we obtain:

$$\frac{\partial h}{\partial p_i}(\mathbf{p}) = h(1_i, \mathbf{p}) - h(0_i, \mathbf{p}).$$

From now on we assume that (N, π) is a fixed coherent system, and all definitions and properties are to be thought in this system even though we do not specify their dependency on it. Also, p denote an arbitrary reliability vector with $\mathbf{p} \in (0, 1)^n$, and i an arbitrary component in N.

Definition 2.1. Birnbaum Importance Measures

a) Birnbaum RIM: $I_i^B(\mathbf{p}) = \dfrac{\partial h}{\partial p_i}(\mathbf{p})$

b) Birnbaum SIM: $I_i^{Bs} = I_i^B(\mathbf{1/2})$ where $\mathbf{1/2} = (1/2, \ldots, 1/2)$

3. THE REPRESENTATIVENESS IMPORTANCE MEASURES

In this section we start by defining the Representativeness reliability importance measure (RIM). The aim of introducing it is to have a measure of the degree of "agreement" between the performance of a given component and the performance of the whole system.

Definition 3.1. Representativeness Importance Measures

a) Representativeness RIM: $I_i^R(\mathbf{p}) = p_i h(1_i, \mathbf{p}) + (1 - p_i)(1 - h(0_i, \mathbf{p}))$

b) Representativeness SIM: $I_i^{Rs} = I_i^R(\mathbf{1/2})$ where $\mathbf{1/2} = (1/2, \ldots, 1/2)$

The first basic properties of the Representativeness RIM are established in the following proposition:

Proposition 3.1.

$i)$ $0 \leq min\{p_i, 1 - p_i\} \leq I_i^R(\mathbf{p}) \leq 1.$

$ii)$ $I_i^R(\mathbf{p}) \geq I_i^B(\mathbf{p}).$

$iii)$ $I_i^R(\mathbf{p}) = (1 - p_i)[1 - 2h(0_i, \mathbf{p})] + h(\mathbf{p}).$
$I_i^R(\mathbf{p}) = p_i[2h(1_i, \mathbf{p}) - 1] + 1 - h(\mathbf{p}).$

$iv)$ $I_i^R(\mathbf{p})$ *is a linear function of* p_i.

$v)$ $I_i^R(\mathbf{p})$ *is non–decreasing in* p_i *if and only if* $h\left((1/2)_i, \mathbf{p}\right) \geq 1/2.$

Let now focus our attention in the Representativeness SIM.

Proposition 3.2. For any component $i \in N$ it is

$$I_i^{Rs} = 1/2 + 1/2\ I_i^{Bs}$$

Observe that we can have $I_i^R(\mathbf{p}) > 1/2 + 1/2\ I_i^B(\mathbf{p})$ and $I_i^R(\mathbf{p}) < 1/2 + 1/2\ I_i^B(\mathbf{p})$ in a system, for the same components *i, j*, depending only on the reliability vector $\mathbf{p}$, as we show in the following example.

Example 3.1. Consider a *3*-out-of-*4* system. If we consider all components having the same reliability p then the Birnbaum RIM and the Representativeness RIM of all components are the same (thus we omit here the arbitrary subscript i) and their respective values are:

$$I^B(p) = 3p^2(1-p) \quad \textit{and} \quad I^R(p) = p^3(2-p) + 1 - p.$$

Now, if $p = 0.6$ *it is* $I^B(0.6) = 0.432$ *and* $I^R(0.6) = 0.7024$ *so that* $I^R(0.6) < 1/2 + 1/2\ I^B(0.6)$,
but for $p = 0.4$ *it is* $I^B(0.4) = 0.288$ *and* $I^R(0.4) = 0.7024$ *so that* $I^R(0.4) > 1/2 + 1/2\ I^B(0.4)$.

In the next theorem we prove that $\mathbf{p} = 1/2$ is the only reliability vector for which the linear relationship stated in Proposition 3.2 is valid between Birnbaum RIM and Representativeness RIM:

Theorem 3.1

$$I_i^R(\mathbf{p}) = 1/2 + 1/2\ I_i^B(\mathbf{p})$$
$$\textit{for any system } \pi, \textit{ and for any component } i$$
$$\Updownarrow$$
$$\mathbf{p} = \mathbf{1/2}$$

Theorem 3.1 proves that the Representativeness RIM and the Birnbaum RIM are not only conceptually different but also analytically independent, as there is no way to derive one concept from the other, and only for the particular case $\mathbf{p} = (1/2, \ldots, 1/2)$ there is an affine relationship between the two measures.

4. COMPARING THE IMPORTANCE OF TWO COMPONENTS

The Representativeness RIM and the Birnbaum RIM rank components of a series (parallel) system in the same way. But this is not true in general, as is shown in the following example.

Example 4.1. Consider the 3–component system given by the reliability function: $h(\mathbf{p}) = p_3 + p_1p_2 - p_1p_2p_3$.

If we take, for instance, $\mathbf{p}$ = (0.6, 0.9, 0.7), then components 1 and 2 verify:

$$\begin{array}{lcl} I_1^R(\mathbf{p}) = 0.702 & < & 0.822 = I_2^R(\mathbf{p}) \\ I_1^B(\mathbf{p}) = 0.27 & > & 0.18 = I_2^B(\mathbf{p}) \end{array}$$

and, thus, the two measures rank them in an opposite order.

It is well known the behavior of the Birnbaum RIM on a series system (the most reliable component has the smallest Birnbaum RIM) and on a parallel system (the most reliable component has the largest Birnbaum RIM), which happens to be the same of the Representativeness RIM in both kinds of systems. Similar properties can be stated for systems which have a component in series (parallel) with the rest of the system, and we prove that they are verified also for the Representativeness RIM. These properties are proved in the next proposition:

Proposition 4.1

a) If a component i is in series with the rest of the system then, for any $j = i$:

$$p_i \geq p_j \quad \Rightarrow \quad I_i^B(\mathbf{p}) \geq I_j^B(\mathbf{p}) \quad \textit{and} \quad I_i^R(\mathbf{p}) \geq I_j^R(\mathbf{p})$$

b) If a component i is in parallel with the rest of the system then, for any $j = i$:

$$p_i \geq p_j \quad \Rightarrow \quad I_i^B(\mathbf{p}) \geq I_j^B(\mathbf{p}) \quad \textit{and} \quad I_i^R(\mathbf{p}) \geq I_j^R(\mathbf{p})$$

Definition 4.1. Node criticality relation

- $i \geq j \leftrightarrow h(1_i, 0_j, \mathbf{p}) - h(0_i, 1_j, \mathbf{p}) \geq 0$ for any $\mathbf{p}$.
- $i \geq j \leftrightarrow i \geq j$ and $h(1_i, 0_j, \mathbf{p}) - h(0_i, 1_j, \mathbf{p}) > 0$ for some $\mathbf{p}$.

Proposition 4.2

- If $i \geq j$ and $p_i = p_j$ then $I_i^R(\mathbf{p}) \geq I_j^R(\mathbf{p})$.
- If $i \geq j$ and $p_i = p_j$ then $I_i^R(\mathbf{p}) \geq I_j^R(\mathbf{p})$.

Next result has a consequence which clarifies the similarities and differences between these two measures.

Theorem 4.1. Let i, j be different components in N .

$i)$ *If* $p_i = p_j$ *then*
$\left[\, I_i^R(\mathbf{p}) > I_j^R(\mathbf{p}) \;\Leftrightarrow\; I_i^B(\mathbf{p}) > I_j^B(\mathbf{p}) \,\right]$ *for any system* π.

$ii)$ *If* $N = \{i, j\}$ *then*
$\left[\, I_i^R(\mathbf{p}) > I_j^R(\mathbf{p}) \;\Leftrightarrow\; I_i^B(\mathbf{p}) > I_j^B(\mathbf{p}) \,\right]$ *for any system* π.

$iii)$ *If* $p_i \neq p_j$ *and there is a component* $k \neq i, j$ *with* $p_k \neq 1/2$, *then we can find a system* π *such that*
$\left[I_i^R(\mathbf{p}) - I_j^R(\mathbf{p})\right]\left[I_i^B(\mathbf{p}) - I_j^B(\mathbf{p})\right] < 0.$

$iv)$ *If* $p_i \neq p_j$ *and for any component* $k \neq i, j$ *it is* $p_k = 1/2$, *then we can find a system* π *such that*
$I_i^B(\mathbf{p}) \neq I_j^B(\mathbf{p})$ *and* $I_i^R(\mathbf{p}) = I_j^R(\mathbf{p})$.

Corollary 4.1. Let $n \geq 3$. Then,

$$\begin{array}{c} \left[I_i^B(\mathbf{p}) > I_j^B(\mathbf{p}) \;\Leftrightarrow\; I_i^R(\mathbf{p}) > I_j^R(\mathbf{p})\right] \\ \text{for any system } \pi \quad \text{and for any } \; i, j \in N \\ \Updownarrow \\ \text{all components of } \mathbf{p} \text{ coincide.} \end{array}$$

Corollary 4.1 states that, if all components of N have the same reliability then, the Representativeness RIM and Birnbaum RIM rank them in the same way, independently of the system taken, but this is the only case in which this property occurs. Thus, under equal reliability of components both measures are qualitatively similar but in any other case the rankings they produce are different. As the two measures order components in different ways for arbitrary component reliabilities we should take into account both measures when analyzing a system. The Representativeness RIM reflects the degree of concordance between the component and the system, or its likelihood of replicating the system performance. In some way it measures how sensitive is a component in the system, and therefore how small changes in this component can cause the loss of this sensitivity. We do think that designers should pay attention to this measure and that it should be taken into account in the near future.

Representativeness measure is the probability that the failure and functioning of the component coincide with system failure and functioning taking into account all possible situations and to analyze the possible rankings that can be obtained.

ACKNOWLEDGMENTS

Research partially supported by Grant MTM2012-34426/FEDER "del Ministerio de Economía y Competitividad".

REFERENCES

Kuo, W., and Zhu, X. (2012). *Importance Measures in Reliability, Risk, and Optimization*. UK: Wiley.

Birnbaum, Z. (1969). "On the Importance of Different Components in a Multicomponent System". *Multivariate Analysis*, vol 2. New York, USA: Academic Press.

39. UNCERTAINTY ANALYSIS IN THE IS-LM MODEL

J. Gabarro and M. Serna
Universitat Politècnica de Catalunya, CS Dept. and ALBCOM Research Group, Spain
{gabarro, mjserna}@cs.upc.edu

ABSTRACT

We introduce a model for uncertainty in the IS-LM linear macroeconomic model with exogenous parameters. An uncertainty profile $\mathcal{U}$ is a short and macroscopic description of an stressed situation. We use $\mathcal{U}$ to define a strategic game where two agents, the angel and the daemon, act selfishly and have different goals. Those Nash equilibria provide the stable strategies in stressed situations, giving a natural estimation of risk. We apply this analysis to a linear version of the IS-LM model and analyse the structure of the Nash equilibria in some particular games.

1. INTRODUCTION

Knight (1921) made a distinction between risk and uncertainty (see also Akerlof, Schiller (2009), Chapter 11): "Risk refers to something that can be measured by mathematical probabilities. In contrast, uncertainty refers to something that cannot be measured because there are no objective standards to express probabilities." We explore a way to transform uncertainty into risk by modelling stressed uncertain situations by an strategic situation among two agents and use game theory to analyse those models. We propose a simplified way to model uncertain situations through an uncertainty profile $\mathcal{U}$. A profile U gives a short and macroscopic description of the potential stress of an economic system, together with the a description of an strategic situation where two agents, the angel a and the daemon d have opposite

goals. Uncertainty profiles were introduced in Gabarro *et al.* (2014) to analyse Web orchestrations under stress in uncertain situations.

We are interested here in extending the approach to the InvestmentSavingsLiquidity Money (IS-LM), introduced in Hicks (1937) (see also Baldani *et al.* (2007)) providing a way to express, in equilibrium, the national income and the interest rate as a a function of several exogenous parameters. Our motivation comes from the fact that one of the limitations in the IS-LM model is the lack of risk. Hicks (1980-1981) wrote that "there is no sense in liquidity, unless expectations are uncertain".

You can imagine a as the government and d as the people, the unions or the crude reality, those agents can act by stressing some of the exogenous parameters of the system in a positive or negative way. This point of view is far from being unique, in southern Europe many citizens believe just the opposite. Thus by considering different uncertainty profiles we can model different situations. This give us a possible algorithmic interpretation of the Keynesian animal spirits[1] in macroeconomics: animal spirits (a and d) are players in a strategic game.

We consider a linear approximation of the IS-LM model, present the global framework of analysis and analyse the Nash equilibria of the resulting $\mathfrak{a}$-$\mathfrak{d}$ for some simple uncertainty profiles. In the studied cases, both $\mathfrak{a}$ and $\mathfrak{d}$ have limited capability to act over the exogenous parameters. The paper is structured as follows. In Section 2 we present the IS-LM model. In Section 3 we model the behaviour of the whole set of exogenous parameters under stress. In 4 we introduce uncertainty profiles and the associated $\mathfrak{a}$-$\mathfrak{d}$ games tailored to the linear IS-LM model and analyse the Nash equilibria of some cases. In particular We prove that adding uncertainty to any fiscal policies generates strategic situations in which there is always a a dominant strategy equilibria. Finally, in Section 5 we raise some lines for future research.

[1] "Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits - a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities" (Keynes (1936) pages 161-162). Look also at Akerlof, Schiller (2009), Part One.

2. THE INVESTMENTSAVINGS-LIQUIDITYMONEY (IS-LM)

We recall here the definition of the IS-LM model, which deals with equilibrium states and is given by two equations. The first equation, the IS (InvestmentSavings) line, describes the equilibrium points in the market of goods and services. The second equation, the LM (LiquidityMoney) line, represents the equilibrium points in the money market.

The IS line, $Y = C\,(Y - T) + I\,(r) + G$, gives an accounting identity where, Y is the national income, Y, r is the interest rate, the remaining parameters are the sum of the annual rates of spending: C by consumers, $I\,(r)$ by the investors (as a function of the interest rate), and G by government G.

The LM line, $M/P = L(r,\ Y\,)$, expresses the fact that the money supply M/P, where M is the money and P the price level, is given by the liquidity preference $L(r,\ Y\,)$ which is a function of the national income Y and the interest rate r (see Keynes (1936) pag 166).

An equilibrium point $(Y,\ r)$ is a pair solving the system. This pair $(Y,\ r)$ gives the point where both markets are on mutual equilibrium.

We consider a linear approximation of the IS-LM model taking $C\,(Y - T) = a + b(Y - T)$, $I\,(r) = c - dr$ and $L(r,\ Y\,) = eY - f\,r$. Those expressions leads to a set of exogenous parameters $\varepsilon = \{a, b, c, d, e, f, T\,, G, M, P\,\}$. The set of endogenous variables $\{Y, r\}$ can be expressed, in equilibrium, as a function of the parameters. The parameters and their constraints are the following:

autonomous consumption	$0 < a$	interest sensitivity for real money	$0 < \mathrm{f}$
marginal propensity to consume	$0 < b < 1$	taxes	$0 < \mathrm{T}$
exogenous investment	$0 < c$	exogenous government spending	$0 < \mathrm{G}$
interest sensitivity	$0 < d$	price index	$0 < \mathrm{P}$
income sensitivity for real money	$0 < e$		

The IS line $Y = Y\,(r)$ is obtained from $Y = C\,(Y - T\,) + I\,(r) + G = a + b(Y - T\,) + c - dr + G$. The LM line $r = r(Y\,)$ is obtained from $M/P = eY - f\,r$. Thus $Y = (a + c + G - bT - dr)/(1 - b)$

and $r = (eY - M/P)/f$. Solving the linear system $Y = Y(r), r = r(Y)$ we obtain the following equilibrium point (Y, r) where $g = (1 - b)f + de$:

$$Y = \frac{f}{g}(a + c + G - bT) + \frac{d}{g}\frac{M}{P} \text{ and } r = \frac{e}{g}(a + c + G - bT) - \frac{(1-b)}{g}\frac{M}{P}$$

As (Y, r) depends on ε, when needed we write $(Y(\varepsilon), r(\varepsilon))$.

Example 2.1. Consider the following valuation or the set ε:

a	b	c	d	e	f	T	G	M	P
200	3/4	200	25	1	100	100	100	1000	2

We get, $Y = 1700 - 100r$, $r = Y/100 + 5$ and solving $Y = 1100$, $r = 6$.

3. STRESS MODEL FOR THE IS-LM MODEL

We assume the existence of two adversary agents called the angel a and the daemon d. Those agents act at the same time attempting to stress the model by altering the values of some of the exogenous parameters of the linear IS-LM model. A stress model for ε is a tuple of pairs $S = ((\delta_a(e), \delta_d(e)))_{e \in \varepsilon}$ describing the perturbation that can be applied to the parameter. The perturbation values are real numbers, so they can be either positive or negative.

Example 3.1. Let us give an example of a stress model S:

agent	a	b	c, d, e, f	T	G	M	P
a	0	+1/20	0	0	+50	0	0
δ	0	0	0	+50	–25	0	+1

The angel a has the (potential) capability to act upon the parameters $\{b, \mathrm{G}\}$. The marginal propensity to consume could be increased from 3/4 to 4/5. This is modelled by $\delta_a(\mathrm{b}) = 1/20$. The government spending G could be increased by $\delta_a(G) = 50$. For any other $e \in \varepsilon \setminus \{\mathrm{b}, \mathrm{G}\}$, the agent a has no possibility to act upon and therefore

$\delta_a(e)$ = 0. The daemon $\mathfrak{d}$ has the following (potential) capability to act upon some of the parameters in {P, T , G}. The price of money could increase by $\delta_d(P)$ = 1. Taxes could increase $\delta_d(T)$ = 50. The G spending could decrease by $\delta_d(G)$ = –25. For any e ∈ $\varepsilon \setminus$ {P, T , G}, $\delta_d(e)$ = 0.

Let us define the transformation of the exogenous parametres under stress model S, when $\mathfrak{a}$ has decided to exert stress capabilities on $a \subseteq \varepsilon$ and $\mathfrak{d}$ has decided to exert stress capabilities on $d \subseteq \varepsilon$. For any $e \subseteq \varepsilon$, the stressed version of e under (a, d) denoted as stress $_S(e)[a, d]$ is defined as follows:

$$\text{stress}_S(e)[a,d] = \begin{cases} e & e \notin a \cup d \\ e + \delta_{\mathfrak{a}}(e) & e \in a \setminus d \\ e + \delta_{\mathfrak{d}}(e) & e \in d \setminus a \\ e + \delta_{\mathfrak{a}}(e) + \delta_{\mathfrak{d}}(e) & e \in a \cap d \end{cases}$$

Finally the equilibrium point is (Y (stress$S$ $(e)[a, d]$), r (stress$S$ $(e)[a, d]$). When S is clear from the context, and we want to emphasize the role of the choice of parameters (a, d) we note $(Y(a, d), r(a, d))$.

Example 3.2. We continue with Example 2.1 and S given in Example 3.1 for (a, d) = ({b}, {P, G}). Writing the new values as e' = stres$S$(e)[{b}, {P, G}] we note stress$S$ (ε) $[\{b\}, \{P, G\}] = \{a', b', c', d', e', f', T', G', M', P'\}$ and:

agent	choice	a	b	c	d	e	f	T	G	M	P
		200	3/4	200	25	1	100	100	100	1000	2
a	$a = \{b\}$		+1/20								
d	$d = \{P, G\}$								–25		+1
		a'	b'	c'	d'	e'	f'	T'	G'	M'	P'
		200	3/4 + 1/20 = 4/5	200	25	1	100	100	100 – 25 = 75	1000	2+1=3

Consider for instance the computation of b'. As $b \in a \setminus d$ = {b} \ $\{P, G\}$ = {b} we have $b' = b + \delta_a$(b) = 3/4 + 1/20 = 4/5. The obtaiined equilibrium point is $Y(\{b\}, \{P, G\}) = 28700/27 \approx 1062.96$, $r(\{b\}, \{P, G\}) = 197/27 \approx 7.29$.

4. UNCERTAINTY PROFILES AND $\mathfrak{a}$-$\mathfrak{d}$ GAMES IN THE IS-LM MODEL

An a priori (global and macroscopic) view of the IS-LM model E in a stressed environment S with uncertainty is modelled by an uncertainty profile $\mathcal{U}$ intro duced in Gabarro *et al.* (2014). In $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \mathcal{A}, \mathcal{D}, b_{\mathcal{A}}, b_{\mathcal{D}}, u_{\mathcal{A}}, u_{\mathcal{D}} \rangle$, $\mathcal{A}$ and $\mathcal{D}$ are subsets of $\mathcal{E}$ which may be stressed. The analyser has the perception that when an angelic parameter in $\mathcal{A}$ is perturbed this is unlikely to have a serious impact (not malicious behaviour). In contrast when a daemonic parameter in $\mathcal{D}$ is perturbed this may well have catastrophic implications (malicious behaviour). The spread of the stress is modelled by $b_{\mathcal{A}}$ and $b_{\mathcal{D}}$. The effects are measured by a utility functions $u_{\mathcal{A}}$ and $u_{\mathcal{D}}$. Given $\mathcal{U}$ a strategic situation arises when the IS-LM model is subjected to combined angel and daemon actions.

An uncertainty profile is a tuple $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \mathcal{A}, \mathcal{D}, b_{\mathcal{A}}, b_{\mathcal{D}}, u_{\mathcal{A}}, u_{\mathcal{D}} \rangle$, where S is a stress model for $\mathcal{E}$, $\mathcal{A} \cup \mathcal{D} \subseteq \mathcal{E}$ and $b_{\mathcal{A}} \leq \#\mathcal{A}$ and $b_{\mathcal{D}} \leq \#\mathcal{D}$.

A $\mathfrak{a}$-$\mathfrak{d}$ strategic game (the $\mathfrak{a}$-$\mathfrak{d}$ game) associated to the uncertainty profile $\mathcal{U}$ is given by $\Gamma(\mathcal{U}) = \langle \{\mathfrak{a}, \mathfrak{d}\}, A_{\mathfrak{a}}, A_{\mathfrak{d}}, u_{\mathfrak{a}}, u_{\mathfrak{d}} \rangle$ where {a, d} are the angel and the daemon players. The player actions are $A_{\mathfrak{a}} = \{a \subseteq \mathcal{A} \mid \#a = b_{\mathcal{A}}\}$ and $A_{\mathfrak{d}} = \{d \subseteq \mathcal{D} \mid \#d = b_{\mathcal{D}}\}$. The utilitites are $u_{\mathfrak{a}} = u_{\mathcal{A}}$ and $u_{\mathfrak{d}} = u_{\mathcal{D}}$.

Example 4.1. Let us continue with Example 3.2. We assume that having high income is good and high interest rate is bad. Then we obtain the bimatrix game $\Gamma(\langle \mathcal{E}, \mathcal{S}, \{b, G\}, \{P, G, T\}, 1, 2, Y, r \rangle)$:

		$\mathfrak{d}$		
		$\{P, G\}$	$\{P, T\}$	$\{T, G\}$
$\mathfrak{a}$	$\{b\}$	$1062.96\ ,\ 7.29$	$1029.62\ ,\ 6.962$	$1233.33\ ,\ 22/3 \approx 7.33$
	$\{G\}$	$1066.66\ ,\ 22/3 \approx 7.33$	$1041.66\ ,\ 7.08$	$1075\ ,\ 5.75$

The PNE are $(\{G\}, \{P, G\})$ $(\{b\}, \{T, G\})$ with different values for Y and r.

It is well know that any strategic game has a Nash equilibrium so from the $\mathfrak{a}$-$\mathfrak{d}$ game, the stable stressed situations are described by the strategies of the players in the Nash equilibria. In the following we analyze the properties of such equilibria for some particular cases.

A case of fiscal policy under uncertainty. Consider the case where a and d have the capability to act over G and T, that is $A_{\mathfrak{a}} = A_{\mathfrak{d}} = \{\{G\}, \{T\}\}$ and $u_{\mathfrak{a}} = Y$ and $u_{\mathfrak{d}} = r$ (as in Example 4.1). We ask if the addition of uncertainty can generate a unstable situation where each agent is trying to catch the other and no PNE exists. Theorem 4.1 shows a negative answer.

Theorem 4.1. Given $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \{T, G\}, \{T, G\}, 1, 1, Y, r\rangle$ where the stress model S verifies is $\delta_{\mathfrak{a}}(x) = \delta_{\mathfrak{d}}(x) = 0$, *for* $x \notin \{T, G\}$, it holds that $\Gamma(\mathcal{U})$ has always a dominant strategy equilibrium.

Proof Sketch. In $\Gamma(\mathcal{U})$ the sets of actions are $A_{\mathfrak{a}} = A_{\mathfrak{d}} = \{\{T\}, \{G\}\}$ and $u_{\mathfrak{a}}(a, d) = Y(a, d)$, $u_{\mathfrak{d}}(a, d) = r(a, d)$. Defining $\mu_{T,T} = -b\big(\delta_{\mathfrak{a}}(T) + \delta_{\mathfrak{d}}(T)\big)$, $\mu_{T,G} = \delta_{\mathfrak{d}}(G) - b\delta_{\mathfrak{a}}(T)$, $\mu_{G,T} = \delta_{\mathfrak{a}}(G) - b\delta_{\mathfrak{d}}(T)$ and $\mu_{G,G} = \delta_{\mathfrak{a}}(G) + \delta_{\mathfrak{d}}(G)$ game $\Gamma(\mathcal{U})$ is

		$\mathfrak{d}$	
		$\{T\}$	$\{G\}$
$\mathfrak{a}$	$\{T\}$	$Y + (f/g)\mu_{T,T},\ r + (e/g)\mu_{T,T}$	$Y + (f/g)\mu_{T,G},\ r + (e/g)\mu_{T,G}$
	$\{G\}$	$Y + (f/g)\mu_{G,T},\ r + (e/g)\mu_{G,T}$	$Y + (f/g)\mu_{G,G},\ r + (e/g)\mu_{G,G}$

it can be proved that $(\{G\}, \{G\})$ is a dominant strategy equilibrium.

Considers a variation where d acts as before but a can act over a or M.

Theorem 4.2. Let $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \{a, M\}, \{T, G\}, 1, 1, Y, r\rangle$ where S is

agent	a	b, c, d, e, f	T	M	G	P
$\mathfrak{a}$	$\delta_a \geq 0$	*0*	0	$\delta_M \geq 0$	*0*	0
$\mathfrak{d}$	0	0	$\delta_T \geq 0$	0	$\delta_G \leq 0$	0

it holds that $\Gamma(\mathcal{U})$ has always a dominant strategy equilibrium.

Proof Sketch. The following $\mathfrak{a}$-$\mathfrak{d}$ game $\Gamma(\langle \mathcal{E}, \mathcal{S}, \{a, M\}, \{T, G\}, 1, 1, Y, r\rangle)$ has dominant strategy equilibriums:

		$\mathfrak{d}$	
		$\{T\}$	$\{G\}$
$\mathfrak{a}$	$\{a\}$	$u_{\mathfrak{a}} = Y + \frac{f}{g}(\delta_a - b\delta_T)$ $u_{\mathfrak{d}} = r + \frac{e}{g}(\delta_a - b\delta_T)$	$u_{\mathfrak{a}} = Y + \frac{f}{g}(\delta_a + \delta_G)$ $u_{\mathfrak{d}} = r + \frac{e}{g}(\delta_a + \delta_G)$
	$\{M\}$	$u_{\mathfrak{a}} = Y + \frac{1}{g}(\frac{d}{P}\delta_M - fb\delta_T)$, $u_{\mathfrak{d}} = r - \frac{1}{g}((1-b)\frac{\delta_M}{P} + eb\delta_T)$	$u_{\mathfrak{a}} = Y + \frac{1}{g}(\frac{d}{P}\delta_M + f\delta_G)$, $u_{\mathfrak{d}} = r - \frac{1}{g}((1-b)\frac{\delta_M}{P} - e\delta_G)$

A case giving zero-sum games. Given $k > 0$ consider the reduced utility $u(a,d) = Y(a,d) - kr(a,d)$ and let $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \mathcal{A}, \mathcal{D}, b_{\mathcal{A}}, b_{\mathcal{D}}, u \rangle$. Then -$u_d$=$u$ and $\Gamma(\mathcal{U})$ is a zero-sum. By similar techniques we have.

Theorem 4.3. Given $\mathcal{U} = \langle \mathcal{E}, \mathcal{S}, \{T, G\}, \{T, G\}, 1, 1, u \rangle$ being S

agent	a, b, c, d, e, f	T	G	M, P
$\mathfrak{a}$	0	$\delta_T \leq 0$	$\delta_G \geq 0$	0
$\mathfrak{d}$	0	$\delta_T \geq 0$	$\delta_G \leq 0$	0

if $(f - ke)\delta = 0$ thethe $\mathfrak{a}$-$\mathfrak{d}$ game has four PNE otherwise it has only one PNE either in ({T}, {T}) or in ({G}, {G}).

5. FURTHER DEVELOPMENTS

We are working towards understanding the structure of the Nash equilibria of $\mathfrak{a}$-$\mathfrak{d}$ games in more complex situations. Varian (1977) has studied the stability of the IS-LM model. As *(Y (a, d), r(a, d))* can be seen as perturbations of the equilibrium point *(Y, r)* if will be of interest to study the relation of both models. Our preliminary result point out that it seems workable to apply the methods in this paper to others parametrized econometric models.

ACKNOWLEDGMENTS

Supported by TIN2013-46181-C2-1-R (COMMAS) and SGR 2014:1137 (ALBCOM).

REFERENCES

Akerlof, G., and Schiller, R. (2009). *Animal Spirits*. Princeton University Press.

Baldani, J., Bradfield, J., and Turner, R. (2007). *Mathematical Economics*. Thomson.

Gabarro, J., Serna, M., and Stewart, A. (2014). "Analysing Web-Orchestrations under Stress Using Uncertainty Profiles". *The Computer Journal*, 57, 11, 1591-1615.

Hicks, J. (1937). "Mr Keynes and the 'Classics': A Suggested Interpretation". *Econometrica*, 5, 2, 147-157.

Hicks, J. (1980-1981). "IS-LM: An Explanation". *Journal of Post Keynesian Economics*, 3, 139-155.

Keynes, J.M. (1936). *The General Theory of Employment, Interest and Money*. Reprint: Palgrave-Macmillan for the Royal Economic Society, 2007.

Knight, F. (1921). *Risk, uncertainty and Profit. Houghton Mifflin.* e-version http://www.econlib.org/library/Knight/knRUP.html.

Varian, H. (1977). "The Stability of a Disequilibrium IS-LM Model". *The Scandinavian Journal of Economics*, 79, 2, 260-270.

40. DYNAMIC SHAPE ANALYSIS - BEFORE AND AFTER CHEMOTHERAPY

P. Hermann
JKU Linz, IFAS, Austria,
philipp.hermann@aon.at

S.M. Giebel
JKU Linz, IFAS, Austria,
stefan.giebel@gmx.de

J.-P. Schenk
University Hospital of Heidelberg, Div. of Pediatric Radiology, Germany,
jens-peter.schenk@med.uni-heidelberg.de

M. Stehlík
JKU Linz/Universidad Técnica Federico Santa María,
IFAS/Departamento de Matemática, Austria/Chile,
mlnstehlik@gmail.com

ABSTRACT

In this study we describe the change of shape by statistical methods. The database consists on Wilms tumors before and after chemotherapy. Our aim is to develop a statistical tool to evaluate the impact of chemotherapy on tumor shape.

1. INTRODUCTION

Wilms tumor, also named nephroblastoma, is a cancer type which mainly affects the kidney of children. Its name is originated from German physicists

Max Wilms, who has made a first bigger study regarding this cancer type (see Wilms, 1889).

Tumor volume of nephroblastomas in childhood is an important risk factor after preoperative chemotherapy after tumor-nephrectomy in therapy protocols and an important parameter for therapy optimization in chemotherapy after tumor resection and nephrectomy in therapy protocol of renal tumors in childhood. Up until now, the tumor shape can be helpful in radiological diagnosis before therapy, but change of tumor shape is not further evaluated after preoperative chemotherapy. Aim of the study is to find a statistical method to evaluate the change of tumor shape in preoperative chemotherapy.

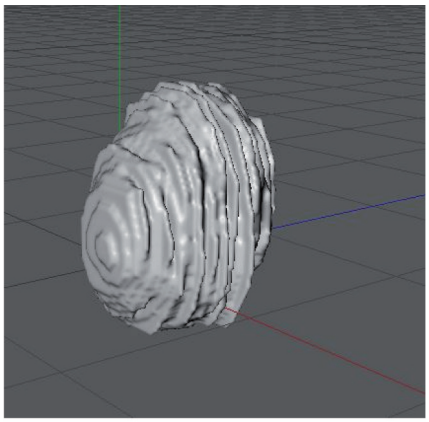


Figure 1. 3D image of a renal tumor.

We propose the use of mathematical procedures to support the applicants in their evaluation of magnetic resonance images (MRI) of renal tumors. We determine key points respectively three dimensional landmarks of retroperitoneal tumors in childhood by using the edges of the platonic body (C60) and test the difference in the objects of the tumor before and after chemotherapy. Objects are centered in order to be comparable before and after therapy. Information of the images of magnetic resonance tomography, especially the renal origin of a tumor and the mass effect with displacement of other organs, is needed for diagnosis (see Schenk *et al.*, 2008). Three dimensional objects are constructed from 2D-MRI images and described by landmarks. Within this study the landmarks are taken as cut points between the

surface of the tumor and the vector of the edge of the platonic body C60. Our study is performed as a clinical trial, where all patients were treated with chemotherapy and the landmarks are measured before and after therapy. Our aim is to show the impact of this therapy on the object of the renal tumor. Figure 1 (taken from Giebel *et al.*, 2012) displays such a 3D image of a renal tumor.

Radiological diagnosis is the most important fact to define for therapy. Therefore, researchers have a huge interest to find good markers which enable a good differentiation of tumors. Hence, proper therapy should be chosen and the development of the tumor respectively the shape of it over time should be observed (see Giebel *et al.*, 2012).

Statistical shape analysis is the keyword in this sense, which is used to measure, to describe and to compare the shapes between objects. In general, the shape of an object is defined as the total information of the corresponding object, data set or image. This shape is invariant regarding to translations, rotations or other isotropic rescalings. Special methods allow to study the shape of objects although the location, rotation or scale of the objects is removed like in this study (see Giebel *et al.*, 2012).

Giebel *et al.* (2010) show that by comparing the mean shape of objects, especially in the field of renal tumors, differentiation between kinds of renal tumors is possible. For this classification neither assumptions regarding the distribution nor any specification of the sample size need to be done.

2. IMPACT OF CHEMOTHERAPY

2.1. Data analysis

For this study, 10 patients were observed before and after chemotherapy. Therefore, it can be compared to a pre and post clinical trial to investigate the impact of chemotherapy on the objects of the renal tumor. In earlier studies differentiation on basis of landmarks between different types of tumors has been conducted. Differences between these groups could be found e.g. in their distributional behavior (see Giebel *et al.*, 2014).

therapy	min	median	mean	max
before	10245	17356	26269	58887
after	3027	19457	44007	231059

Table 1. Descriptive statistics of measured points.

Coordinate - therapy	min	median	mean	max	mean of sds
x-before	-66.94	0.55	-0.02	72.74	7.09
x-after	-53.64	0.36	0.18	56.36	7.14
y-before	-84.79	-0.18	-1.13	77.65	8.00
y-after	-66.74	-0.36	-0.30	60.96	7.67
z-before	-75.72	-0.40	0.15	63.28	7.30
z-after	-53.64	0.36	0.18	56.36	7.14

Table 2. Descriptive overview of the coordinates by therapy status.

This data set consists of the patient number, status whether therapy has been finished, number of measured points, x- y and z-coordinates for the 60 landmarks (180 variables) and the center for the three landmarks. In total there are 20 observations, whereby each patient was observed twice, i.e. before (status = 1) and after (status = 0) therapy. For allowance of comparison of the size of the object, each landmark was centered according to its center value. In Table 1 some descriptive statistics of the number of measured points are listed. It can be seen that the median as well as the mean number of measured points for defining the tumor is lower before therapy. The platonic body is described by landmarks which are classified on an explorative approach (see Giebel *et al.*, 2010). Therefore, the number of measured points varies between 3027 and 231059 in order to receive a valid shape of the tumor.

Descriptive statistics should give a first indication about the possible impact of the therapy. Table 2 compares the five-point summary and the mean of the standard deviations of every axis and therapy status. A decrease in the mean of the standard deviation can be observed for y- and z-coordinate after therapy. Moreover, minima and maxima are smaller for all coordinates after therapy, which allows the assumption that therapy could lead to a decrease of the volume/shape of the object.

2.2. Density estimation

In order to have a graphical comparison, density estimators were plotted in Figure 2. For this purpose, kernel density estimations were performed in R (R Core Development Team, 2008) with the aid of the function density on basis of a Gaussian kernel. Thereby, within every plot one specific coordinate of the landmarks is represented and its distributional difference before and after therapy is of interest. Obviously, therapy leads to less variance in the distribution for every coordinate, which strengthens the assumption arising from Table 2 before.

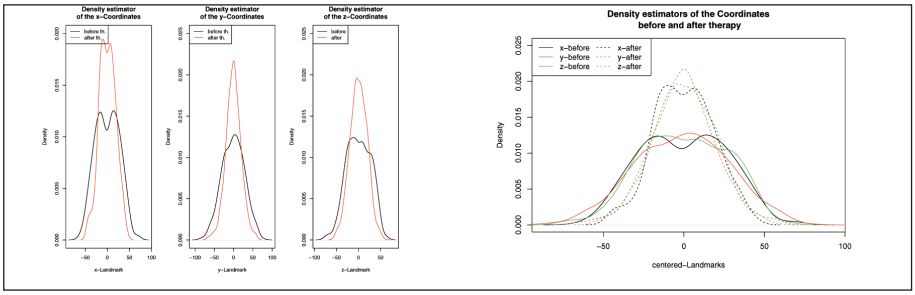


Figure 2. Density estimators for every coordinate (left) and in one figure

Another indication of Figure 2 is that the distribution of the coordinates could be assumed as normal distributed after therapy. Shapiro-test was used for this purpose and yields for y-coordinate before therapy ($p = 0.35$) as well as y- and z-coordinate after therapy ($p = 0.08$ respectively 0.75) in p-values, which do not allow to reject the null-hypothesis of normal distribution. Although normality is still rejected for x-coordinate, its p-value at least increases (from < 0.0001 to 0.02) and a development in direction of normality can be assumed, among others because of the normal shape of the density estimator. Further research on the underlying distributions by usage of mixture distributions, as suggested by e.g., the two modi of x-coordinate after therapy, could obtain information about the distributional behavior after therapy.

2.3. Graphical analysis

Further graphical investigations on the effect of chemotherapy based on histogram and QQ-plot can be done in Figure 3. A slight change in direction of normality is

visible for all coordinates. Especially the histograms show shapes, which are easier comparable to normal distribution after therapy than before.

In order to have a graphical impression on the fit of normal distribution of the data, mean and standard deviation of the data were taken to fit this distribution (see Table 2). Figure 4 shows that usage of corresponding MLEs leads to a better distributional fit of the landmarks for patients after therapy. Especially for y- and z-coordinate normality seems to be a proper underlying distribution. Result of this testing procedure can be that more abnormal tissue leads to larger deviations from normality. Hence, therapy could have a positive impact on removing cancerous tissue.

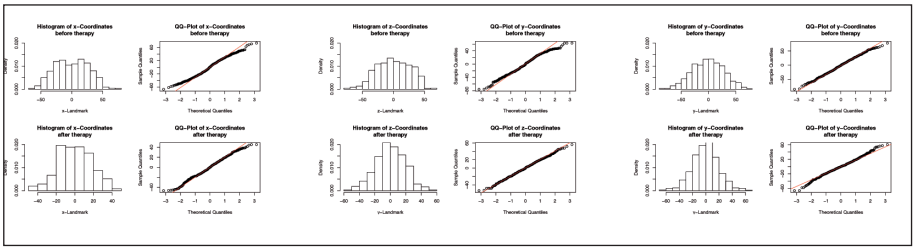


Figure 3. Histogram and QQ-plot for the coordinates, and the therapeutical effect where left: x, middle: y and right: z.

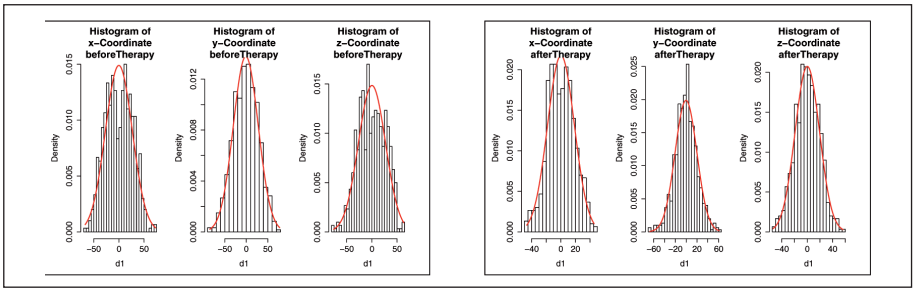


Figure 4. Fit of the landmarks before (left) and after (right) therapy with normal distribution.

As a next step the previous assumption of the impact of the therapy on the volume of the object is analyzed graphically via 3D-scatterplots in Figure 5. Therefore, six of the patients are compared to have a graphical indication of the development of the object of the tumor before and after therapy. Plots of the landmarks before therapy

can be found in the first and third column and the corresponding plots after therapy are displayed on their right hand side. Apparently, the volume/shape of the object decreases for all displayed patients (all axes have the same scale). Yet, this reduction develops in different magnitudes as can be seen by comparing the effect for e.g., patients 1 and 3 to patient 2. However, cancer is known to have fractal properties, wherefore calculations of euclidean volumes are not proper methods in contrast to using fractal volume measures to check for the impact of the therapy on cancer.

3. CONCLUSION

Even in this step of our research on Wilms tumors, we can prove that mathematical procedures can support medical decisions in radiology. In our first study on 2d-shapes the euclidean approach was productive. For computing 3d-shapes, shape analysis had to be combined with a neural network to get an optimal differentiation. Shape analysis, dynamic statistical shape analysis and fractal analysis have a potential to describe the change of a tumor after chemotherapy in a mathematical way, differentiation of therapy relevant tumors groups should be evaluated in further studies.

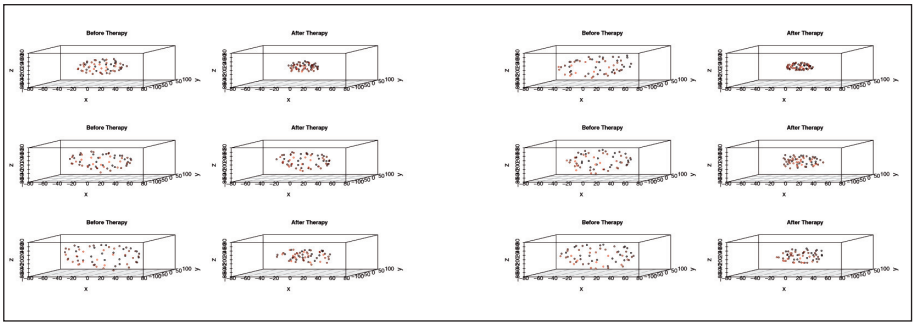


Figure 5. 3D-Scatterplot for six patients before and after therapy.

ACKNOWLEDGMENTS

Milan Stehlík acknowledges Proyecto Interno 2015, REGUL. MAT 12.15.33, Modelación del crecimiento de tejidos con aplicaciones a la investigación del cáncer. Philipp Hermann acknowledges ANR project Desire FWF I 833-N18.

REFERENCES

Giebel, S.M., Hermann, P., Schenk, J.-P., and Stehlík, M. (2014). *On Analytical Methods for Cancer Research*. ICCRA Conference, Springer (in print).

Giebel, S.M., Schlitz, J., and Schenk, J.-P. (2010). "Application of Shape Analysis on Renal Tumors in 3D". *Health Informatics and Bioinformatics (HIBIT)*, 5th International Symposium on 20-22 April 2010.

Giebel, S.M., Schlitz, J., Graf, N., Nourkami, N., Leuschner, I., and Schenk J.-P. (2012). "Application of Shape Analysis on 3D Images - MRI of Renal Tumors". *JIRSS*, 11(2), pp 131-146.

R Development Core Team (2008). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schenk, J.P., Graf, N., Günther, P., Ley, S., Göppl, M., Kulozik, A., Rohrschneider, W.K., and Tröger, J. (2008). "Role of MRI in the Management of Patients with Nephroblastoma". *Eur Radiol*. 2008 Apr, 18 (4), 683-691.

Wilms, M. (1899). *Die Mischgeschwülste der Niere*. Leipzig: Verlag von Arthur Georgi, 1-90.

41. PREDICTING WIND SPEEDS FOR ENERGY RISK: WEIBULL DISTRIBUTION AND NEURAL NETWORK METHODS AS ALTERNATIVE AND COMPLEMENTARY STRATEGIES

S.M. Giebel
University, Linz, Austria,
stefan.giebel@enamec.de

M. Rainer
ENAMEC Institut and FZ Risk Management,
Univ. Würzburg, Germany,
martin.rainer@enamec.de

ABSTRACT

Prediction of weather factors is becoming increasing important for risk management for both, the renewable energy sector and the conventional energy sector. In this paper we consider wind speed in particular.

Wind speed can be modelled as a stochastic process with certain timedependent features (parameters) of its distribution. In order to capture some of its extreme value features, the Weibull distribution is quite common.

As an alternative to stochastic models, neural networks are frequently applied. While there are less preconditions (distributions, dependencies etc.), there are also a lot of disadvantages of conventional neural network methods: Black-Box procedure, overfit, dependency on initial parameters and the presumption that every explaining variable can be dependent on each other. As we have shown in our previous research (Giebel, Rainer and Aydin, 2013), there are a lot of possibilities to combine Neural networks with stochastic models (for example also the estimation of distribution parameters for the residuals).

In the current work, we present a possibility to estimate directly the parameters of the Weibull distribution by a suitable neural network. We compare the potential of forecasting with a linear regression and a direct neural network on three different samples.

1. THE CHALLENGE OF STRUCTURE(D) LEARNING WITH STOCHASTIC PROCESSES

In a series of recent investigations (Giebel and Rainer, 2009; Giebel and Rainer, 2011; Giebel, Rainer and Aydin, 2013) progress has been made, in order to adapt learning algorithms to the specific structure of some typical stochastic processes.

Let the state variable X be given by a simple CAR(n) process — the continuous analogue of the time-series AR(n) — described by the following system of SDEs:

$$\begin{aligned} dX_q &= X_{q+1}(t)\,dt \quad q = 1, \ldots, n-1 \\ dX_n &= -\left(\sum_{q=1}^{n} \alpha_q(t) X_q(t)\right) dt + \sigma(t) dI(t) \end{aligned}$$

In the special case of $n = 1$, the systems reduces to a (generalized) Orstein-Uhlenbeck process with mean-reversion $\alpha_1(t)$ and volatility $\sigma(t)$.

An observable Y depending on X can then be modeled as

$$Y(t, X(t)) = g\left(\Lambda(t) + X_1(t)\right)$$

with any given semimartingale driver I.

In order to yield a more explicit form of the dynamics of Y, in terms of an SDE for dY, the general Itô formula for semimartingales should be applied to $Y = g(X)$.

Typical applications of the considered processes included models for financial markets, energy markets, energy resources, and climate factors.

2. STOCHASTIC MODELS

We assume that a common seasonality $\eta(t)$ for positive and negative jump sizes can be factorized as

$$J^{\pm}(t) = \eta(t) J^{\pm} \tag{1}$$

For the deseasonalized jump sizes $J^{\pm}$, let us assume a generalized gamma distribution:

$$F_{J,b,p}{}^{\pm} := \frac{b^{\pm}}{\mu_J^{\pm}\Gamma\left(\frac{b}{p}\right)} \int_0^z \left(\frac{z}{\mu_J^{\pm}}\right)^{b^{\pm}-1} e^{-(\frac{z}{\mu_J^{\pm}})^{p^{\pm}}} dz \tag{2}$$

Weibull distribution

As important special cases, this includes the gamma distribution for $p = 1$, the Weibull distributions for $p = b$, and the log-normal distributions in a limit $\frac{b}{p} \to \infty$.

$$F_{J,b}{}^{\pm} := \frac{b^{\pm}}{\mu_J^{\pm}} \int_0^z \left(\frac{z}{\mu_J^{\pm}}\right)^{b^{\pm}-1} e^{-(\frac{z}{\mu_J^{\pm}})^{b^{\pm}}} dz \tag{3}$$

Moreover, for $b = 1$ it reduces to the common exponential distribution, which is our particular choice for this study. On the other hand, for $b > 1$ the probability density vanishes for vanishing jump sizes. Parameter b is sometimes called the Weibull modul. If small jumps are very rare, a Weibull module of $b > 1$ may be desirable. In this case, a Weibull or a generalized gamma distribution provide an alternative to the log-normal distribution. This is actually the case for wind speeds, where the shape parameter under normal conditions differs only little from $b = 2$ (the Rayleigh distribution).

3. A MULTI-OU MODEL FOR WIND SPEEDS

Wind speeds may well be described by the following stochastic spike-jump-diffusion model. It is composed of multiple OU processes with independent semimartingale drivers, namely

$$\begin{aligned} S(t) &= g\left(\Lambda(t) + X(t) + Y^{+}(t) + Y^{-}(t)\right) \\ dX &= -\alpha X dt + \sigma(t)(dW + dL) \\ dY^{\pm} &= -\beta^{\pm} Y^{\pm} dt \pm \eta(t) J^{\pm} dN^{\pm}(t) \end{aligned} \tag{4}$$

where W is a continuous Brownian motion and L a pure jump process, e.g. a compound Poisson process. J ± is the random variable for the heights of positive and negative spiky jumps, N ± are the Poisson subordinators with spike intensity λ±(t), which is assumed to be a continuous function of time t.

For the purpose of wind speeds the geometric choice $g(x) := e^x$ is suitable. The distributions of positive and negative spiky jump heights are modeled commonly by two symmetric exponential distributions, ie,

$$F_{J^\pm} := \pm\frac{1}{\mu_{J^\pm}}\int_0^x e^{\mp(\frac{z}{\mu_{J^\pm}})}dz = 1 - e^{\mp(\frac{x}{\mu_{J^\pm}})} \tag{5}$$

respectively, where $\mu_J \pm$ is the expected size of the positive and negative spiky jumps. The corresponding semimartingale (compensator) measure is

$$\ell(dt, dz)^\pm \;=\; dt\frac{\lambda^\pm(t)}{\mu_J^\pm}e^{\mp(\frac{z}{\mu_J^\pm})}dz \tag{6}$$

4. MULTI-LAYER PERCEPTRONS

A multi-layer perceptron network performs a training step by modifying the weights of all input layers. The learning mechanism the weights is determined by the target distance measure

$$E \;=\; \frac{1}{2}\sum_{i=1}^{n}(y^i - \bar{y}^i)^2 \quad .$$

The weights of both layers are changed according to the steepest descent, i.e.

$$\nabla_{w_i}E \;:=\; \frac{\partial E}{\partial w_i} \tag{7}$$

$$\nabla_{u_j}E \;:=\; \frac{\partial E}{\partial u_j} \tag{8}$$

With a learning rate α, which should be adapted to the data, the weights are changed as follows:

$$w_i^{new} = w_i^{old} - \alpha \cdot \nabla_{w_i} E \tag{9}$$

$$u_j^{new} = u_j^{old} - \alpha \cdot \nabla_{u_j} E \tag{10}$$

5. NEURAL NETWORK & STOCHASTIC MODEL

We use a 2-layer neural network, where we compute variance and mean via weighted input variables. The first estimate of the mean is

$$\mu_0 = \frac{1}{n} \sum_{i=1}^{n} w_i x_i \quad ,$$

and for the variance it is

$$\sigma_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (w_i x_i - \mu_0)^2 \quad .$$

The neural network is then trained by adjusting the weights w_i of the first layer, and the weights u_j of the second layer, according to the respective sensitivities (7) and (8) of the error function.

6. NEURAL PARAMETER CALIBRATION

In this section we demonstrate the dynamical calibration for a log-normal stochastic process based on mean and variance only.

In the first layer μ and σ are determined on the basis of weighted values $w_i y_i$ of the time series of y-values.

$$\hat{\mu}_y := \frac{1}{\sum_{i=1}^{n} dt_i} \sum_{i=1}^{n} w_i dy_i \tag{11}$$

$$\hat{\sigma}^2 := \frac{1}{\sum_{i=1}^{n} dt_i} \sum_{i=1}^{n} w_i dt_i \left[\frac{1}{dt_i} \left(\frac{dS_i}{S_{i-1}} \right)^2 \right] \tag{12}$$

In the second layer a new value is determined with

$$\hat{y}_{i+1} \quad = \quad y_i + (u_1 \hat{\mu}_y + u_2 \hat{\sigma}) dt_i \quad . \tag{13}$$

Including this value in the regression the process for determining μ and σ in the first layer is repeated. In every iteration the weights in the first and second layers are changed according to the steepest descent.

7. NUMERICAL RESULTS

As an example, we test our structural stochastic and neuro-stochastic models with the daily mean wind speed in 0.1 m/s measured at Zugspitze.

In every prediction step, previous values are used to predict the next one. The models were implemented in MATLAB and Mathematica, respectively. In all figures the predicted values are compared to the real values (faint lines). The models can be compared on the basis of their coefficients of determination, R^2.

In the first approach the daily wind speeds are predicted by a stochastic model and the residuals by a neural model. In the second approach we use directly a neural modell 1. Using the stochastic model and multi-layer neural networks, respectively, we find for Zugspitze $R^2_{\text{sm}} = 0.25$ and $R^2_{\text{nn}} = 0.12$. Hence, around 12 to 25 per cent of the fluctuations could be explained by at least one of the compared models.

In any case a considerable fraction of wind speed fluctuations is still explained by the neural network, although the residuals are supposed to contain information only about spikes and autoregression, features which are not modeled explicitly by the neuro-stochastic model.

The p-value in all examples is $p < 0.001$ is considerably smaller than the significance level of 0.05, hence the correlation between reality and prediction is highly significant. As expected, only a low level of positive correlation ranging between 0.346 and 0.436 is achieved in the case of neural networks. This range improves as much as 0.42-0.50 in stochastic approach. By using only our combination between Weibul-distribution and a neural network, we get folllowing figure.

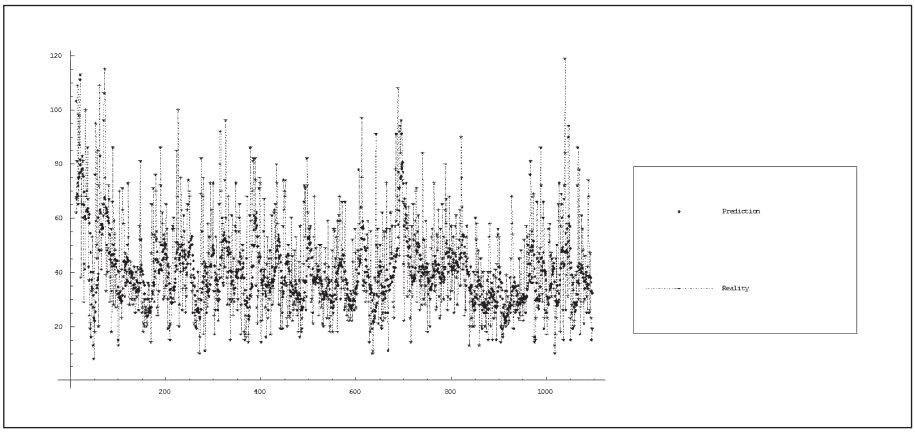


Figure 1. Wind speed (in units of 0.1 m/s) at Zugspitze: daily predictions l (bold black spots) between 1.1.2010-31.12.2010 with 2 years of previous history.

Direct prediction of daily wind speed by a combination of Weibul-distribution and a neural network leads us to a similiar result of $R^2 = 0,24$. That means, that the structural stochastic model has no advantage in comparison to our approach. Our approach could be easily improved by a second neural network to predict the residuals.

Considering a target time horizon of 2 years, good agreement of the predicted values with the reality is achieved. Calibrating the model-parameters with neural weights adapting continuously to the history, the model learns to capture the optimal trend according to movements of the past years.

8. CONCLUSION AND OUTLOOK

In our combined model of neural networks and stochastic processes, the neural transformation processes adapt in such a manner that, from a continuously updated history of fixed length the network is continuously learning the process parameters. In comparison to traditional calibration, our neural methods is taking into account the historical process in a more detailed and dynamical manner, as it is shown in (Giebel and Rainer, 2013). As a result, the parameters of the stochastic process can be better calibrated than with traditional methods. As our examples demonstrate:

Even a relatively simple stochastic process, in combination together with a smart neural network learning continuously the right parameters, the model yields satisfactory predictions. This demon- strates the efficiency of our combined stochastic-neural approach.

The research is continued to apply our neural network approach also to more demanding stochastic processes, as they may be required for estimations of different climate factors, or energy resources, or energy demand and prices, such as for gas or electricity.

In several such cases, different phases of normal and spiky modes of volatility have to be taken into account. This can be done on the one hand by regime switching models, on the other hand by working with more advanced semi-martingale processes including jumps, also with time-dependent frequency.

Neural networks calibration may improve stochastic models. In the light of their potential to capture all kinds of generalized (linear and non-linear) "trends", neural networks can give us an idea, how much of structural, i.e. "trend" information we loose, if we do not improve our results, especially in the case of renewable energy. Unraveling of hidden structural information could be economically important, e.g. for energy storage and reduction of non-renewable forms of energy. Furthermore it could show up hidden weather trends. In order to apply neural networks, we have to account a priori for a correlation between all measurements. These correlations have to be checked carefully. Furthermore the validity of our model has to be demonstrated, by applying the model on real unknown data, in particular since generally conclusive measures for model validity are not easily available either.

In the broader context of learning algorithms (aka learning machines), the multilayer perceptron (MLP) applied in the second model above is unusual, by the fact that we use tailored activation functions, describing well defined parameters of a stochastic model. Hence every neuron in the intermediate layer has a clear interpretation. This is very different from the common data mining versions of the multilayer perceptron (MLP) and its generalizations, the support vector machines (SVMs). In the standard MLP applied for some activation function, the intermediate layers are usually a hidden black box. Similarly, the feature space of a nonlinear SVM

typically remains without interpretation, and the right choice of the kernel function is rarely evident. In this aspect our neuro-stochastic model provides an interesting new path for the design of learning algorithms, with a feature space clearly defined by the parameters of the stochastic model, the kernel functions (which here are just the activation functions) designed directly for learning these model parameters. Opening this black box might give a new direction to further research in statistical learning.

REFERENCES

Giebel, S.M., and Rainer, M. (2009). "Forecasting Financial Asset Processes: Stochastic Dynamics via Learning Neural Networks". *Bull. Soc. Sciences Médicales Luxembourg*, 10 (1), 91-107.

Giebel, S.M., and Rainer, M. (2011). "Stochastic Processes Adapted by Neural Networks with Application to Climate, Energy, and Finance". *Applied Mathematics and Computation*, 218, 1003-1007.

Giebel, S.M., Rainer, M., and Aydin, N. (2013). "Simulation and Prediction of Wind Speeds: A Neural Network for Weibull". *Journal of Iranian Statistical Society*, vol. 12, no. 2, 293-320

Giebel, S.M., and Rainer, M. (2013). "Neural Network Calibrated Stochastic Processes: Forecasting Financial Assets". *Centr. Europ. J. Operations Research*, 21 (2), 277-293.

Maqsood, I., Khan, M.R., and Abraham, A. (2004). "An Ensemble of Neural Networks for Weather Forecasting". *Neural Computing & Applications,* June, volume 13, issue 2, 112-122.

42. EXTENSIONS AND APPLICATIONS OF COMPARE: WEB PLATFORM TO SELECT THE PRIMARY ENDPOINT IN A RANDOMIZED CLINICAL TRIAL

M. Gómez-Mateu and G. Gómez Melis
Universitat Politècnica de Catalunya, Department of Statistics and Operations Research, Spain,
moises.gomez.mateu@upc.edu

ABSTRACT

We present CompARE, a web platform that computes how more efficient is to use a composite endpoint instead of one of its components. CompARE is based on the Asymptotic Relative Efficiency (ARE) method developed by Gómez and Lagakos. Users interact with CompARE through HTML form pages and no knowledge of R is needed. A list of candidate endpoints, anticipatable probabilities and hazard ratios are required to use CompARE. Results are shown immediately through plots, text and tables. We present advanced capabilities of CompARE such as assigning probabilities and hazard ratios to combinations of several endpoints (relevant or additional).

1. INTRODUCTION

Conclusions from randomized clinical trials (RCTs) rely on its primary endpoint chosen at the design stage of the study. In order to provide clinical evidence related to the primary objective of the trial the selection of the primary endpoint is of outmost importance.

Composite endpoints (CEs) consisting of the union of two or more outcomes are often used as the primary endpoint in RCTs. For instance, in time-to-event studies, time to MACE is a CE generally defined as the union of cardiovascular death,

reinfarction, target vessel revascularization and stroke. Pros and cons on the use of CEs have been extensively discussed in the literature. From a medical perspective, the use of CEs allows to combine multiple measurements, avoiding the need to choose a single outcome as primary endpoint. However, some cautions have to be considered when defining a composite endpoint. A CE should only be used if the individual components are clinically meaningful and of similar importance to the patient, the expected effects on each component are similar based on biological plausibility, and the clinically more important components do not affect negatively (Ferreira-González *et al.*, 2007). Otherwise, the use of the CE is questionable and may lead to misleading conclusions. From a statistical point of view, the rationale for considering a CE, other than multiplicity, is the likely reduction of the sample size needed to achieve a desired statistical power for a given significance level, as a result of increasing the event rate to attain a predefined treatment effect. Nevertheless, it has been discussed and shown in Gómez and Lagakos (2013), that adding inappropriate components to the relevant endpoint might actually lead to a loss of power to detect the true treatment differences.

2. THE ARE METHOD TO CHOOSE THE PRIMARY ENDPOINT

Gómez and Lagakos (2013) develop a statistical methodology that helps to decide between using a Relevant endpoint ε_R instead of a composite endpoint, consisting of the union of ε_R plus another additional endpoint ε_A, to evaluate the effect of a treatment. Their strategy is based on the value of the asymptotic relative efficiency (ARE).

The ARE relates the efficiency of using the logrank test Z_R based on ε_R versus the efficiency of the logrank test Z_* based on the CE. It can be shown that Z_R and Z_* are asymptotically $N(0, 1)$ under the null hypothesis of no treatment effect and asymptotically $N(\mu_R, 1)$ and $\mathrm{N}(\mu_*, 1)$, respectively, under a sequence of contiguous alternatives to the null hypothesis. Under these conditions, $\mathrm{ARE}(Z_*, Z_R) = (\mu / \mu)^2$. The composite endpoint will be advisable to use as primary endpoint whenever the $\mathrm{ARE}(Z_*, Z_R) > 1$. Otherwise, ε_R will be considered the best choice.

We assume that: i) the end-of-study censoring is the only non-informative censoring cause for both groups; ii) the hazard ratios HR_R and HR_A for ε_R and ε_A, respectively,

are constant; iii) marginal Weibull distributions for the time T_R and T_A to ε_R and ε_A, respectively; and iv) Frank's copula binding the marginals (T_R, T_A) through Speraman correlation ρ. Under these assumptions, when the additional endpoint ε_A does not include death, the expression of the ARE is given by:

$$\mathrm{ARE}(Z_*, Z_R) = \frac{\left(\int_0^1 \log\left\{\lambda_*^{(1)}(t)/\lambda_*^{(0)}(t)\right\} f_*^{(0)}(t)dt\right)^2}{\left(\log\left\{\mathrm{HR_R}\right\}\right)^2 \cdot\, p_{*} \cdot\, p_R},$$

where p_R and p_* are the probabilities of observing ε_R and ε_* in control group, respectively; and $\lambda_*^{(0)}(t)$ and $\lambda_*^{(1)}(t)$ are the hazard functions of the times $\mathrm{T}_*^{(0)}$ and $\mathrm{T}_*^{(1)}$ to the composite endpoint for each group, respectively. The ARE expression in either censoring case is expressed in terms of the following interpretable parameters:

- Probabilities p_R, p_A of observing ε_R, ε_A, respectively, in control group,
- relative treatment effects given by the hazard ratios HR_R, HR_A, and
- Spearman's correlation coefficient ρ between T_R and T_A.

The ARE as ratio of sample sizes

If we were to test H_0 versus H_a with two different test statistics S_n and T_m , the asymptotic relative efficiency would be defined as the ratio m/n, where n and m are the required sample sizes for S_n and T_m, respectively, to attain the same power for a given significance level.

In our setup we have two different set of hypotheses: H_0 versus H_a testing the treatment effect on ε_R and H_0^* versus H_a^* testing the treatment effect on 0 a ε_R. Gómez and Lagakos prove that the ARE can be interpreted as n_R/n_*, where n_R and n_* stand for the required sample sizes needed when using the Relevant endpoint or the composite endpoint, respectively.

3. CompARE: A TOOL TO CHOOSE THE PRIMARY ENDPOINT IN A RCT

With the aim of making the ARE method widely applicable to the scientific community, a free web-based platform called *CompARE* has been created.

CompARE provides a tool to compute the ARE among several subsets of relevant components. It is of great help when planning a clinical trial since it quantifies how efficient is a relevant subset of outcomes respect to a larger subset. CompARE is a friendly tool, based on the free software Tiki Wiki CMS/Goupware. Although it is internally programmed in R, users do not need knowledge of R, neither to have R installed in their computer. By means of web-page forms, users can easily introduce the required information, step by step. This information is saved and executed by the system, returning ad-hoc results depending on each case.

ARE values are calculated from the anticipated information of the probabilities and hazard ratios p_R, p_A, HR_R and HR_A by means of an input grid. Moderate correlations ρ and exponential distributions are considered by default. Other advanced options are available such as Weibull distributions with decreasing or increasing hazard rates or different correlations.

Results from CompARE are shown by means of summary tables and plots. Outputs combining several HR_A together with different correlation values are of great help to understand the role that the Additional endpoint plays (see Figure 1). Additionally, survivals and hazard ratios for ε_R, ε_A and CE are depicted. A history table saves each result to compare previous analyses. Moreover, conclusions and recommendations are given in written form as an aid. CompARE is currently accessible as a beta version on the following website (http://composite.upc.edu/CompARE).

4. APPLICATIONS TO THE CARDIOVASCULAR RESEARCH AREA

The ARE method has been applied to the cardiovascular research area (Gómez, Gómez-Mateu and Dafni, 2014), where a set of general recommendations are reported from RCTs published in 2008. General guidelines are provided based on the frequencies of observation and treatment effects of the Relevant and the Additional

endpoints. In the following subsection we illustrate some of the CompARE capabilities by means of a real RCT.

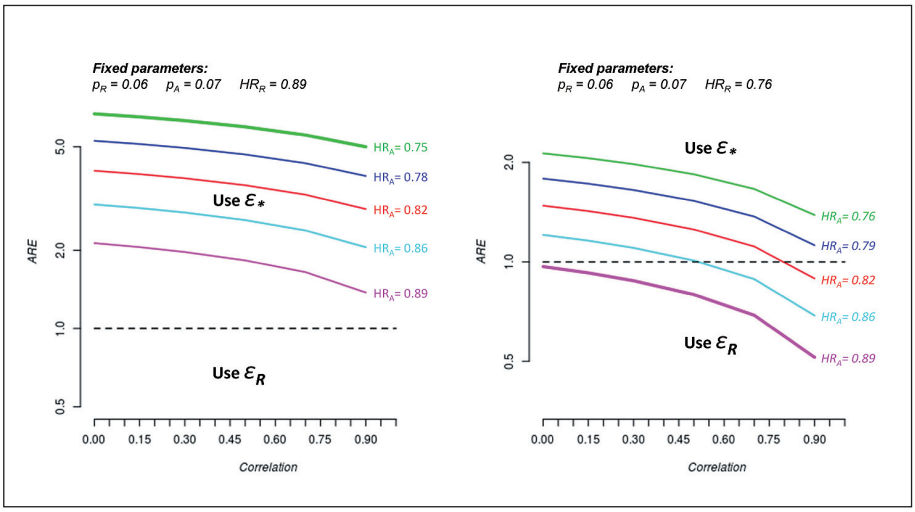


Figure 1. ARE values combining different values of the hazard ratio HR_A of the Additional endpoint ε_A and correlation between times to the Relevant endpoint ε_R and ε_A. Fixed parameters: probabilities p_R and pA of observing ε_R and ε_A, respectively, in control group; and hazard ratio HR_R of ε_R.

A case study: The LIFE trial

The Losartan Intervention For Endpoint reduction in hypertension study (LIFE) trial (Dahlöf *et al.*, 2002) was performed to test the efficacy of Losartan-based antihypertensive treatment in patients with hypertension. The primary composite endpoint (CE) was composed by cardiovascular death, myocardial infarction and stroke. Cardiovascular death + myocardial infarction are considered the most clinically important components (Sankoh *et al.*, 2014), and hence we refer to them as the Relevant endpoint ε_R, and stroke as the Additional endpoint ε_A. Significant results were achieved using the CE, consequence of the significant effect on stroke (see Table 1).

Endpoint	Type	Control treatment Probability (n)	Hazard ratio (CI, 95%)	p-value
Cardiovasc. mortality	ε_{r1}	0.05 (234)	0.89 (0.73 – 1.07)	0.206
Myocardial infarction	ε_{r2}	0.04 (188)	1.07 (0.88 – 1.31)	0.491
Stroke	ε_A	0.07 (309)	0.75 (0.63 – 0.89)	0.001
ε_{r1} U ε_{r1} U ε_A	Composite	0.13 (588)	0.87 (0.77 – 0.98)	0.021

Table 1. Summary of the results from the LIFE trial. ε_{r1} and ε_{r2} stand for Relevant endpoint component 1 and 2 respectively; ε_A stands for Additional endpoint; and CI stands for the confidence interval.

We present an ARE study based on two different situations. In both we will assume a probability of observing ε_R and ε_A in control group of p_R = 0.06 and p_A = 0.07, respectively, as feasible anticipated values of this study.

If we anticipate relative treatment effects as reported in the LIFE study, that is HR_R = 0.89 for the relevant endpoint and HR_A = 0.75 for the additional endpoint, we would have recommended to use the composite endpoint, as they did. Indeed, as Figure 1 (left) shows, the ARE value is always greater than 1 for every possible correlation between ε_R and ε_A.

However, if the expected relative treatment effect on the relevant endpoint would have been stronger, for instance HR_R = 0.76, then the CE would have not been always advised. As shown in Figure 1 (right), we would conclude that using ε_R as primary endpoint would be the best recommendation whenever the relative treatment effect of the Losartan intervention on stroke was larger than 0.9. On the other hand, adding stroke to the primary endpoint would be advisable if the relative treatment effect was at most 0.8. For the situations where the relative treatment effect on stroke were expected to be between 0.8 and 0.9 special attention should be paid at the correlation between cardiovascular death + myocardial infarction and stroke.

5. PRACTICAL ISSUES TO ASSIGN ANTICIPATABLE COMBINED PROBABILITIES AND HAZARD RATIOS

The ARE method is based on the assumption that, even when the Relevant endpoint consist of several components, both combined probability and hazard ratio can be

anticipated by researchers (analogously, for the Additional endpoint). In some occasions trialists might anticipate these probabilities and hazard ratios for each one of the components. In what follows, we describe how to compute the combined probability and hazard ratio and discuss how it is implemented in CompARE.

Following Bahadur (1961), the combined probability p_R can be computed in terms of the probability components p_{R1} and p_{R2} of observing the relevant components ε_{R1} and ε_{R2} in the control group, respectively, for a given correlation δr between endpoints (more than two components can also be considered). Concerning the combined HR_R, the exact value is in terms not only of the hazard ratios HR_{r1} and HR_{r2} of ε_{R1} and ε_{R2}, respectively, but also depends on the probabilities p_{r1} and p_{r2}, the marginal laws of each component, the correlation and the joint distribution between components. Furthermore, HR_R does not have to be constant even though HR_{r1} and HR_{r1} are constant.

CompARE implements, as advanced options, the possibility of assigning a value for p_R and for HR_R. In particular, whenever the correlation between ε_{R1} and ε_{R2} is unknown, the user can choose between any value within the following bounds: max(p_{r1}, p_{r2}) and $p_{R1} + p_{R2}$ (see Figure 2). For the combined HR_R a set of possible values is as well proposed.

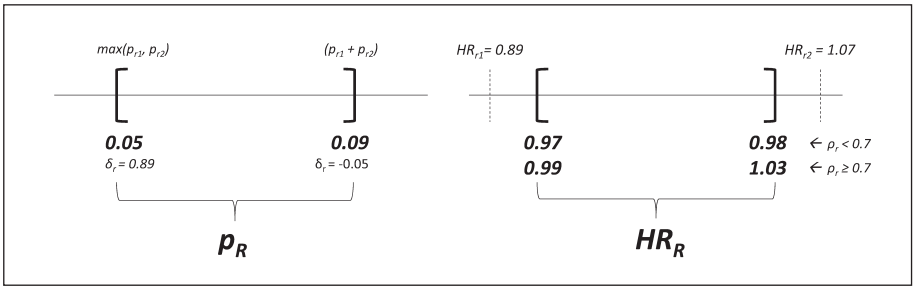


Figure 2. Boundaries for the combined probability p_R and hazard ratio HR_R in terms of Pearson's correlation δr and Spearman's correlation ρ_r, respectively. p_{r1} and p_{r2} stand for the marginal probability of observing the Relevant component ε_{r1} and ε_{r2} in control group, respectively. HR_{r1} and HR_{r2} stands for the marginal hazard ratio of each component. Component parameter values are taken from the LIFE study.

6. CONCLUDING REMARKS

This paper addresses the ARE method as an intuitive and interpretable way of comparing the efficiency of several endpoints, candidates for the primary endpoint of a RCT. We present as well CompARE, a web platform that performs all the ARE computations, which can be freely accessed and is friendly to use. Ongoing research include several extensions of CompARE :

1. Sample size computation based on the anticipatable parameters.

2. Combined probabilities and hazard ratios.

3. Computations when both $\mathcal{E}_R$ and $\mathcal{E}_A$ include death.

4. Different copulas other than Frank's.

ACKNOWLEDGMENTS

This research is partially supported by Grant MTM2012-38067-C02-01 from the Ministerio de Economía y Competitividad and by Grant 2014 SGR 464 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya.

REFERENCES

Bahadur, R.R. (1961). "A Representation of the Joint Distribution of Responses to n Dichotomous Items". *Stanford University Press*, 158-168.

Dahlöf, B., Devereux, R.B., Kjeldsen, S.E., Julius, S., Beevers, G., de Faire, U., Fyhrquist, F., Ibsen, H., Kristiansson, K., Lederballe-Pedersen, O., Lindholm, L.H., Nieminen, M.S., Omvik, P., Oparil, S., Wedel, H., and LIFE Study Group (2002). "Cardiovascular Morbidity and Mortality in the Losartan Intervention For Endpoint Reduction in Hypertension Study (LIFE): A Randomised Trial Against Atenolol". *Lancet*, 359: 995-1003.

Ferreira-González, I., Permanyer-Miraldaa G., Busse J.W., Bryant D.M., Montori V.M, Alonso-Coello P., Walter S.D, and Guyatt G.H. (2007). "Methodologic Discussions for Using and Interpreting Composite Endpoints are Limited, but still Identify Major Concerns". *J Clin Epidemiol*, 60, 651-657.

Gómez, G., and Gómez-Mateu, M. (2014). "The Asymptotic Relative Efficiency and the Ratio of Sample Sizes when Testing two Different Null Hypotheses". *SORT*, 38, 73-88.

Gómez, G., Gómez-Mateu, M., and Dafni, U. (2014). "Informed Choice of Composite End Points in Cardiovascular Trials". *Circulation*, 7: 170-178.

Gómez, G., and Lagakos, S.W. (2013). "Statistical Considerations when Using a Composite Endpoint for Comparing Treatment Groups". *Statistics in Medicine*, 32, 719-738.

Gómez-Mateu, M., and Gómez, G. (Last date of access: January 13, 2015). http://composite.upc.edu/CompARE.

Sankoh, A.J., Lia, H., D'Agostino, R.B. (2014). "Use of Composite Endpoints in Clinical Trials". *Statistics in Medicine,* 33: 4709-4714.

43. SELECTION OF THE PRIMARY ENDPOINT IN A CLINICAL TRIAL: COMPOSITE BINARY ENDPOINTS VERSUS SINGLE OUTCOMES

G. Gómez
Universitat Politècnica de Catalunya, Statistics and Operations Research Department, Barcelona, Spain,
lupe.gomez@upc.edu

A. Ayestaran
Institut Català d'Oncologia, IDIBELL, Hospitalet, Barcelona, Spain,
aintzane_aie@hotmail.com

ABSTRACT

A composite binary endpoint consisting of the combination of several binary outcomes is often used in the design of a two-arm randomized clinical trial. We present a method to elucidate between a composite binary endpoint and one of its relevant components. This approach is based on the asymptotic relative efficiency (ARE) between two statistical tests to compare the effect of the treatment between two groups. The ARE is expressed in terms of clinical interpretable parameters, such as, the anticipated values of the probability of exhibiting the relevant endpoint and the additional endpoint, the correlation between both endpoints and the anticipated treatment effect in the relevant and additional endpoint. A case study illustrates its use.

1. INTRODUCTION

Clinical trials are a key research tool for advancing medical knowledge and patient care. A clinical trial is a prospective biomedical research study of human subjects designed to answer specific questions about interventions (vaccines, drugs, treatments,

devices, etc). Clinical trials are used to determine whether new biomedical interventions are safe, efficacious, and effective. The endpoint or target variable that measures the clinical evidence is a key decision that has to be defined in the protocol of the study.

In a clinical trial the primary outcome has to be clinically relevant, free of bias, feasible to measure and sensitive to the study intervention. In many clinical trials an endpoint fulfiling all these conditions is often difficult to find. In these ocassions the combination of various outcomes into a unique endpoint could be a good solution (Ferreira-González *et al.*, 2007).

Several papers have approached the planning of a randomized clinical trial in the context of time-to-event analysis when the primary endpoint is composite, although only few have attempted to quantify the gained efficiency by incorporating additional components. Gómez and Lagakos (2013) proposed

a method to quantify this addition by using Pitman's asymptotic relative efficiency. Suppose that the intervention could affect two possible endpoints: ϵ_1 and ϵ_2 and define ϵ_* the composite endpoint as the union of ϵ_1 and ϵ_2. Let T_1, T_2 and $T_* = \min\{T_1, T_2\}$ be the times until events ϵ_1, ϵ_2 and ϵ_*. Assuming that a logrank test is used to compare the two groups by means of either the relevant endpoint T_1 or the composite endpoint T_*, the asymptotic relative efficiency (ARE) between both tests measures the gained efficiency obtained by incorporating an additional component and using, as primary, the composite endpoint. The criterion used is to choose the composite endpoint ϵ_* whenever $\mathrm{ARE} > 1$ and keep the relevant endpoint ϵ_1 as the primary endpoint for the trial if $\mathrm{ARE} \leq 1$.

The ARE is a very convenient tool to be used at the planning stage of a clinical trial. Although its mathematical expression is cumbersome and depends on several parameters, decisions can be made based on a short number of anticipated values, as it is often the case when computing the required sample sized in a clinical trial. The ARE can be interpreted as the reciprocal ratio of the sample sizes required for the two logrank test based on T_1 and on T_* to attain the same power at the same significance level (Gómez and Gómez-Mateu, 2014). Thus, the ARE can be used not only as a tool to decide whether or not to use the CE instead of the relevant, but also to

compute the needed sample size if the CE is to be used as primary endpoint. A web platform CompARE facilitates its use.

This paper approaches the planning of clinical trials with composite binary outcomes. Special attention, as Rauch and Beyersmann (2013) point out, has to be given to the proper analysis and interpretation trying to achieve a compromise between the use of the composite endpoint and the significance of each of their components.

2. NOTATION AND MAIN ASSUMPTIONS

We assume a two-arm randomized clinical trial to demonstrate the effectiveness of a treatment. To this end, two different binary endpoints which might provide evidence on the effect of the treatment are identified: a relevant endpoint, ϵ_1 and an additional one, ϵ_2. The composite endpoint $\epsilon_* = \epsilon_1 \cup \epsilon_1$ is defined as the event that occurs whenever one of the two endpoints is observed.

For a sample of $N = N_0 + N_1$ subjects, N_0 are allocated to group 0 with probability π. For each subject, a relevant (k = 1) and an additional (k = 2) endpoints are recorded. Let Y_{ijk} = 1 whenever the j^{th} subject in the ith group exhibits response k and let p_{ik} represent the probability that an individual in treatment group i exhibits response k. Denote by $Y_{ik} = \sum_{j=1}^{N_i} Y_{ijk}$ the total $j=1$ number of individuals that exhibit the kth response (k = 1, 2, j = 1, . . . , N_i, i = 0, 1).

Denote by Y_{ij*} (i = 0, 1, $j = 1, \ldots, N_i$) the indicator for the composite endpoint ϵ_*, for the jth subject in the ith group, that is, $Y_{ij*} = 1$ if $Y_{ij1} + Y_{ij2} \geq 1$ 2 N_i and $Y_{ij3} = 0$ if $Y_{ij1} + Y_{ij2} = 0$, and denote by $Y_{i*} = \sum_{j=1}^{N_i} Y_{ij*}$ the total number j=1 of individuals that exhibit at least one of the two different types of response (relevant or additional). The probability p_{i*} that an individual in group i exhibits at least one of the two different types of response can be derived from the joint distribution of (Y_{ij1}, Y_{ij2}). The joint distribution between any pair of binary random variables is uniquely determined by the probabilities (p_{i1}, p_{i2}) and the correlation $p_i = Corr(Y_{ij1}, Y_{ij2})$ (Bahadur, 1961) and its joint probability function is given by

$$P\left[Y_{ij1} = y_{ij1}, Y_{ij2} = y_{ij2}\right] = \prod_{k=1}^{2} \left(p_{ik}^{y_{ijk}} \cdot q_{ik}^{1-y_{ijk}}\right) \left(1 + \rho_i \cdot z_{ij1} \cdot z_{ij2}\right) \quad (1)$$

where $z_{ijk} = \frac{y_{ijk} - p_{ik}}{\sqrt{p_{ik} q_{ik}}}$ and $q_{ik} = 1 - p_{ik}$. Following (1), the probability p_{i*} can be expressed in terms of the marginal probabilities (p_{i1}, p_{i2}) together with the correlation ρ_i for the ith group: $p_{i*} = 1 - P[Y_{ij*} = 0] = 1 - q_{i1} q_{i2} - \rho_i \sqrt{p_{i1} p_{i2} q_{i1} q_{i2}}$.

3. TESTING THE TREATMENT EFFECT BY MEANS OF ϵ_1 AND ϵ_*

The potential treatment effect under investigation can be established either via the relevant endpoint ϵ_1 or by means of the composite endpoint, ϵ_*, using each one a different set of hypotheses. The null hypothesis of no treatment effect based on ϵ_1 is given by $H_{01} : p_{01} = p_{11}$, or equivalently, in terms of the odds ratio, $H_{01} : \mathrm{OR}_1 = \frac{p_{11}/1-p_{11}}{p_{01}/1-p_{01}} = 1$. Analogously, the null hypothesis of no treatment effect based on ϵ_* is given by $H_{0*} : p_{0*} = p_{1*}$, or equivalently, by $H_{0*} : \mathrm{OR}_* = \frac{p_{1*}/1-p_{1*}}{p_{0*}/1-p_{0*}} = 1$.

The complex relationship between the marginal odds ratios and the composite odds ratio is shown in (2):

$$\mathrm{OR}_* = \frac{(\mathrm{O}_{01}\mathrm{OR}_1 + 1)(\mathrm{O}_{02}\mathrm{OR}_2 + 1) - 1 - \rho_1\sqrt{\mathrm{O}_{01}\mathrm{OR}_1\mathrm{O}_{02}\mathrm{OR}_2}}{\frac{1}{q_{01}q_{02}} - 1 - \rho_0\sqrt{\mathrm{O}_{01}\mathrm{O}_{02}}} \cdot \frac{1 + \rho_0\sqrt{\mathrm{O}_{01}\mathrm{O}_{02}}}{1 + \rho_1\sqrt{\mathrm{O}_{01}\mathrm{OR}_1\mathrm{O}_{02}\mathrm{OR}_2}} \quad (2)$$

where $\mathrm{O}_{01} = \frac{p_{01}}{1-p_{01}}$, $\mathrm{O}_{02} = \frac{p_{02}}{1-p_{02}}$ and $\mathrm{OR}_2 = \frac{p_{12}/1-p_{12}}{p_{02}/1-p_{02}}$.

This relationship clearly shows that H_{0*} is not equivalent to H_{01} and, therefore, when we are conducting either test, we are not measuring the same treatment effect. Only if H_{01} is true, OR_2 = 1 and the correlation between the relevant and the additional endpoints is the same in both groups, that is, $\rho_0 = \rho_1$, then OR_* = 1, although, even in this case, the reciprocal is not necessarily true. Caution is needed when interpreting the findings based on a CE.

Theorem: Two sample proportion tests are considered for both H_{01} and H_{0*}. Denoting by $\tilde{p}_1 = \frac{Y_{01}+Y_{11}}{N_0+N_1}$ and $\tilde{p}_* = \frac{Y_{0*}+Y_{1*}}{N_0+N_1}$ the pooled sample proportions, the tests are the following:

1. $U_1 = \sqrt{\frac{N_0+N_1}{N_0 N_1 \bar{p}_1 \bar{q}_1}}\,(N_0 Y_{11} - N_1 Y_{01})$

2. $U_* = \sqrt{\frac{N_0+N_1}{N_0 N_1 \bar{p}_* \bar{q}_*}}\,(N_0 Y_{1*} - N_1 Y_{0*})$

3. Both tests are asymptotically $N(0, 1)$, under $H_{01} : p_{01} = p_{11}$ and $H_{0*} : p_{0*} = p_{1*}$, respectively.

4. The efficiency of U_1 and U_* is studied by examining their large sample behaviour when the null hypothesis does not hold. We view the null as fixed, and consider a sequence of contiguous alternatives to the null that approaches the null as $n \to \infty$. The sequence of alternatives can be chosen as $\mathrm{OR}_1 = \exp\{h_1/\sqrt{N}\} < 1$ and $\mathrm{OR}_* = \exp\{h_*/\sqrt{N}\} < 1$, respectively. It can be shown that both U_1 and U_* are, under these alternatives, asymptotically normal with variance 1 and means $\mu_1 = \sqrt{\pi(1-\pi)p_{01}q_{01}}$ and $\mu_* = \sqrt{\pi(1-\pi)p_{0*}q_{0*}}$, respectively.

4. ASYMPTOTIC RELATIVE EFFICIENCY OF U_* VERSUS U_1

To assess the difference in efficiency between using U_1 based on ϵ_1 or U_* based on ϵ_*, we use the asymptotic relative efficiency (ARE) of U_* versus U_1. The ARE is given by the square of the ratio of the non-centrality means μ_1 and μ_*, that is,

$$ARE(U_*, U_1) = \left(\frac{\mu_*}{\mu_1}\right)^2 = \frac{(\log(\mathrm{OR}_*))^2}{(\log(\mathrm{OR}_1))^2}\frac{p_{0*}q_{0*}}{p_{01}q_{01}}, \tag{3}$$

where $p_{0*} = 1 - q_{01}q_{02} - \rho_0\sqrt{p_{01}p_{02}q_{01}q_{02}}$.

Hence, the computation of the ARE is based on the following easily anticipatable parameters: (i) the frequencies p_{01} and p_{02} of observing endpoints ε_1 and ε_2 in treatment group 0; (ii) the relative treatment effects on ε_1 and ε_2 given by the odd ratios OR_1 and OR_2; and (iii) the correlations, ρ_0, ρ_1 between the relevant endpoint ε_1 and the additional endpoint ε_2 in treatment groups 0 and 1.

To present the results for situations where it is reasonable to assume $\rho_0 = \rho_1 = \rho$, we propose plots as shown in Figure 1. For a given clinical trial and for plausible anticipated

values for, at least, p_{01}, p_{02} and OR_1, the corresponding plot will reveal the range of values of OR_2 and of ρ for which ARE is greater than 1 and the composite endpoint is advisable.

5. ACUTE MIGRAINE STUDY

Ho *et al*. (2008) conducted a randomised, parallel-treatment, placebocontrolled and double-blind clinical trial in Europe and USA from February 2007 to October 2007. 1380 adults with acute migraine diagnosed by International

Headache Society criteria suffering moderate or severe attacks were included in the study. Each subject was randomly assigned to receive telcagepant 150 mg (n=333) or 300 mg (354), zolmitriptan (345), or placebo (348). The primary aim of this study was to confirm the efficacy and safety profile of telcagepant compared with placebo. To this end, the following five outcomes were used as co-primary endpoints: pain freedom, pain relief, or absence of photophobia, phonophobia, or nausea at 2 h after treatment.

For the purpose of this illustration we only consider the comparison of telcagepant 300 mg versus placebo. We shall assume that the relevant endpoint is to suffer migraine pain and the additional is to experiment phonophobia. From the reported values in the study we have that the probability of suffering pain is p_{01} = 0.904, the probability of experimenting phonophobia is p_{02} = 0.632, the anticipated treatment effects are: OR_1 = 0.288 to reduce pain and OR_2 = 0.425 to reduce phonophobia.

Given p_{01} and p_{02}, the correlations ρ_0 and ρ_1 are not free to range over (–1, +1). For each i = 0, 1, ρ_i is within M_i and m_i (Sozu *et al.*, 2010) where

$$M_i = \max\left\{-\sqrt{\frac{p_{i1}p_{i2}}{(1-p_{i1})(1-p_{i2})}}, -\sqrt{\frac{(1-p_{i1})(1-p_{i2})}{(p_{i1})(p_{i2})}}\right\}$$

and

$$m_i = \min\left\{\sqrt{\frac{p_{i1}(1-p_{i2})}{(1-p_{i1})(p_{i2})}}, \sqrt{\frac{(1-p_{i1})(p_{i2})}{p_{i1}(1-p_{i2})}}\right\},$$

and applying these bounds to our parameter values, we have:

$$M_0 = -0.248 \leq \rho 0 \leq 0.427 = m_0 \text{ and } M_1 = -0.709 \leq \rho 1 \leq 0.582 = m_1.$$

We have computed the ARE for the 88 possible combinations of: p_{01} = 0.904, p_{01} = 0.632, OR_1 = 0.288 together with OR_2 = 0.3, 0.4, 0.6 and 11 different correlation values within the above bounds. In 18.18% of these cases we have ARE$\rightarrow$ 1, thus it would be recommended to add the additional endpoint of phonophobia to the relevant endpoint of pain.

We illustrate the behaviour of the ARE by means of Figure 1 where we are assuming that the correlation between migraine pain and phonophobia is the same in both groups: telcagepant 300 and placebo. Since $\rho_0 = \rho_1$, the common correlation ρ between the relevant and the additional endpoint in either group can vary between $\max(M_0, M_1) = -0.248 \leq \rho \leq \min(m_0, m_1) = 0.427$. In this case, we have 21 scenarios, i.e., the combination of OR_2 = 0.3, 0.4, 0.6 and 7 different correlation values included in the restriction of ρ.

In all these scenarios ARE < 1, hence it would always be more efficient to use migraine pain as primary endpoint.

Observe that when the treatmente effect in the additional endpoint increases (for smaller values of OR_2), the ARE also increases, indicating the advantage of using the CE whenever the treatment effect measured in the additional endpoint is stronger.

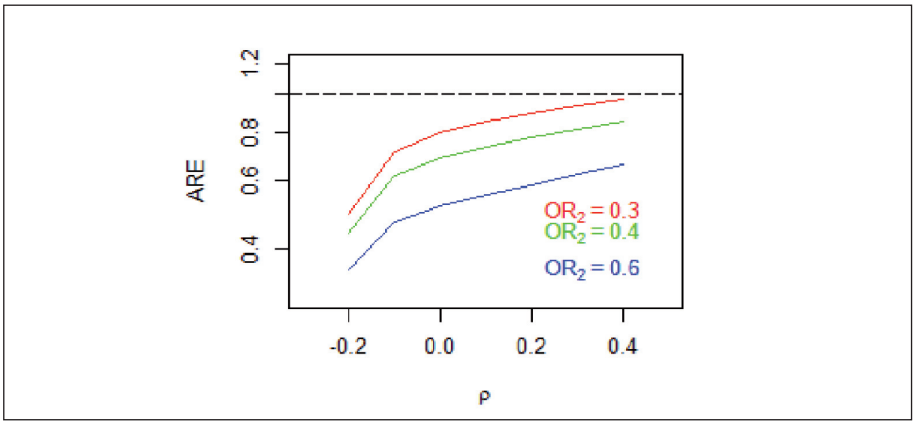


Figure 1. ARE plots for the migraine study, where the probability of suffering migraine pain is p_{01} = 0.904, of experiencing phonophobia is p_{02} = 0.632, treatment effect for migraine pain is OR_1 = 0.288, for phonophobia is $\mathrm{OR}_2 \in \{0.3, 0.4, 0.6.\}$ and for correlations $\rho_0 = \rho_1 = \rho \in \{-0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4\}$.

ACKNOWLEDGMENTS

This work is partially supported by grant MTM2012-38067-C02-01 from the Ministerio de Economía y Competitividad del Gobierno de España and 2014 SGR 464 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya.

We are indebted to the GRASS group (http://grass.upc.edu) for fruitful discussions.

REFERENCES

Bahadur, R.R. (1961). "A Representation of the Joint Distribution of Responses to *n* Dichotomous Items". *Studies in Item Analysis and prediction*. Stanford Univ. Press, 158-168.

Ferreira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D.M., Montori, V.M., Alonso-Coello, P., Walter, S.D., and Guyatt, G.H. (2007). "Methodologic Discussions for Using and Interpreting Composite Endpoints are Limited, but still Identify Major Concerns". *Journal of Clinical Epidemiology*, 60, 651-657.

Gómez, G., and Lagakos, S.W. (2013). "Statistical Considerations when Using a Composite Endpoint for Comparing Treatment Groups". *Statistics in Medicine*, 32, 719-738.

Gómez, G., and Gómez-Mateu, M. (2014). "The Asymptotic Relative Efficiency and the Ratio of Sample Sizes when Testing two Different Null Hypotheses". *SORT*, 38, 73-88.

Gómez-Mateu, M., and Gómez, G. (Last date of access: December 5, 2014). http://composite.upc.edu/CompARE.

Ho, T.W., Ferrari, M.D., Dodick, D.W., Galet, V., Kost, J., Fan, X., Leibensperger, H., Froman, S., Assaid, C., Lines, C., Koppen, H., and Winner P. K. (2008). "Efficacy and Tolerability of MK-0974 (Telcagepant), a New Oral Antagonist of Calcitonin Gene-Related Peptide Receptor, Compared with Zolmitriptan for Acute Migraine: A Randomised, Placebo-Controlled, Parallel-Treatment Trial". *Lancet*, 372, 2115-2123.

Rauch, G., and Beyersmann, J. (2013). "Planning and Evaluating Clinical Trials with Composite Time-to-First-Event Endpoints in a Competing Risk Framework". *Statistics in Medicine*, 32, 3595-3608.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). "Sample Size Determination in Clinical Trials with Multiple co-Primary Binary Endpoints". *Statistics in Medicine*, 29, 2169-2179.

44. THE EXPONENTIAL-RECIPROCAL GENERALIZED INVERSE GAUSSIAN. DISTRIBUTION. A NEW MODEL TO DESCRIBE LOSS DISTRIBUTIONS

E. Gómez-Déniz and F.J. Vázquez-Polo
University of Las Palmas de Gran Canaria, Department of Quantitative Methods and TiDES Institute, Spain, {emilio.gomez-deniz, francisco.vazquezpolo}@ulpgc.es

V. García
University of Cádiz, Dpt. of Statistics and O.R., Spain, victoriano.garcia@uca.es

E. Calderín-Ojeda
University of Melbourne, Dpt. of Economics, Australia, enrique.calderin@unimelb.edu.au

ABSTRACT

A new family of probability distributions with support in (0, ∞) is introduced here. This new family is obtained through the mixture of the exponential distribution with the reciprocal of the generalized inverse Gaussian distribution. We show some important features such as expressions of probability density function, moments, etc. Special attention is paid to the mixture with the reciprocal of the inverse Gaussian distribution, as a particular case of the reciprocal of the generalized inverse Gaussian distribution. This distribution can be considered competitive to the exponential-inverse Gaussian distribution in Frangos and Karlis (2004) and has a more simple formulation than this. The versatility of this family is proven with a numerical example.

1. INTRODUCTION

A new family of probability distributions with support in (0, ∞) is introduced here. This new family is obtained through the mixture of the exponential distribution with the reciprocal of the generalized inverse Gaussian distribution, a distribution which has not been well studied in the statistical literature. We show some important features such as expressions of probability density function, moments and survival function. Special attention is paid to the mixture with the reciprocal of the inverse Gaussian distribution, as a particular case of the reciprocal of the generalized inverse Gaussian distribution, which avoid to use the modified Bessel function of the second kind. This distribution can be considered competitive to the exponential-inverse Gaussian distribution in Frangos and Karlis (2004) having a more simple formulation than this. The versatility of this family is proven with a numerical example. The new model and some of its properties is presented in Section 2. Here, some methods of estimations are also provided. A numerical application is shown in Section 3 and conclusions in the last Section.

2. THE NEW FAMILY OF DISTRIBUTIONS

The reciprocal of the inverse Gaussian distribution is not very common among statisticians and has been little explored in the literature. Furthermore, the reciprocal of the generalized inverse Gaussian distribution has not been encountered by the authors of this work in the statistical literature. It is simple to see that

$$g_{\gamma,\delta,\nu}(z) = \left(\frac{\gamma}{\delta z}\right)^{\nu+1} \frac{\delta}{2\gamma K_{\nu}(\delta\gamma)} \exp\left[-\frac{1}{2}\left(\frac{\gamma^2}{z} + \delta^2 z\right)\right], \tag{1}$$

for $z > 0, \gamma > 0, \delta > 0$ and $K_n(\cdot)$ representing the modified Bessel function of the second kind (Jørgensen, 1982), is the probability density function (pdf) of the reciprocal of a variable distributed according to the generalized inverse Gaussian distribution with parameters $\gamma > 0, \delta > 0$ and $v \in \mathbb{R}$. That is, if Y, $Y > 0$, follows an generalized inverse Gaussian distribution with parameters $\gamma > 0, \delta > 0$ and $v \in \mathbb{R}$, then the random variable $Z = 1/Y$ follows the distribution given in (1). See, for instance, Jørgensen *et al.*

(1991)[1]. From now, when a positive random variable Z follows the pdf given in (1) we will write $Z \sim \mathcal{RGIG}(\gamma, \delta, \nu)$.

Definition 2.1. Let X a continuous random variable with support in $(0, \infty)$ and with probability density function proportional to $f(x) \propto \exp(x/\theta)$, i.e. with the exponential distribution. We say that the random variable X follows an exponential-reciprocal generalized inverse Gaussian distribution if it admits the following stochastic representation.

$$X|\theta \sim \mathcal{E}(\theta),\ \theta \sim \mathcal{RGIG}(\gamma, \delta, \nu),\ \theta \in \Theta = (0, \infty),$$

being $\gamma > 0$, $\delta > 0$ and $\nu \in \mathbb{R}$. From now on, this distribution will be denoted by $X \sim \mathcal{ERGIG}(\gamma, \delta, \nu)$.

The new distribution has mode at zero as it is well–known for mixture of exponentials. By conditioning it is easy to obtain the pdf of the new model, given by

$$f_{\gamma,\delta,\nu}(x) = \left(\frac{\gamma}{\phi(x,\gamma)}\right)^{\nu+1} \frac{\delta K_{\nu+1}(\delta\phi(x,\gamma))}{\gamma K_{\nu}(\gamma\delta)}, \tag{2}$$

for $x > 0$, being $\gamma > 0$, $\delta > 0$, $\nu \in \mathbb{R}$ and $\phi(x,\gamma) = \sqrt{2x + \gamma^2}$.

Proposition 2.1. The mean, variance and survival function of the $\mathcal{ERGIG}$ distribution in (2) are given by

$$E(X) = \frac{\gamma}{\delta}\frac{K_{\nu-1}(\gamma\delta)}{K_{\nu}(\gamma\delta)}, \tag{3}$$

$$var(X) = \left(\frac{\gamma}{\delta}\right)^2 \left\{\frac{K_{\nu-2}(\gamma\delta)}{K_{\nu}(\gamma\delta)} - \left[\frac{K_{\nu-1}(\gamma\delta)}{K_{\nu}(\gamma\delta)}\right]^2\right\}, \tag{4}$$

$$\bar{F}_{\gamma,\delta,\nu}(x) = \left(\frac{\gamma}{\delta}\right)^{\nu} \frac{K_{\nu}(\gamma\phi(x,\gamma))}{K_{\nu}(\gamma\delta)}. \tag{5}$$

[1] Observe that there exist several different parameterizations of the inverse Gaussian distribution (see, for example Chhikara and Folks, 1989; and Seshadri, 1983, 1993).

Proof. The result is obtained easily by compounding, having into account that the mean, second order moment about the origin and survival function of the exponential distribution are given by θ, $2\theta^2$ and $\exp(-x/\theta)$, respectively.

The reciprocal of the inverse Gaussian distribution is obtained as a particular case of (1) when $v = -1/2$. In this case (2) is reduced to

$$f_{\gamma,\delta}(x) = \frac{\delta}{\phi(x,\gamma)} \exp\left[\delta(\gamma - \phi(x,\gamma))\right],$$

a pdf which is reminiscent, but more simple, of the exponential-inverse Gaussian distribution (see Bhattachrya and Kumar, 1986; and Frangos and Karlis, 2004).

In this case, (3), (4) and (5) are reduced to

$$\begin{aligned} E(X) &= \frac{1+\gamma\delta}{\delta^2}, \\ var(X) &= \frac{2+\gamma\delta}{\delta^4}, \\ \bar{F}_{\gamma,\delta}(x) &= \exp\left[\delta(\gamma - \phi(x,\gamma))\right]. \end{aligned}$$

2.1. Parameters estimation

In order to get the estimates of the parameters of the distributions we can consider the method of moments and the maximum likelihood method. This is not deal here. The solutions of the log-likelihood equation can be obtained easily by direct numerical search for the global maximum of the log-likelihood surface. Furthermore the class of generalized reciprocal inverse Gaussian distributions is a natural conjugate class prior for the exponential distribution, as it can be seen in the following result.

Theorem 2.1. Let $X_i|\theta \sim \mathcal{E}(\theta)$, $i = 1, 2, \ldots, n$ be independent and identically distributed random variables. Let us suppose that the parameter θ follows a generalized reciprocal inverse Gaussian prior distribution, $\theta \sim \mathcal{RGIG}(\gamma, \delta, \nu)$. Then, the posterior distribution of parameter θ for the sample $\underline{X} = (X_1, \ldots, X_n)$ is given by $\theta|\underline{X} \sim \mathcal{RGIG}(\tilde{\gamma}, \tilde{\delta}, \tilde{\nu})$, where

$$\tilde{\gamma} = \gamma + 2n\bar{X}, \quad \tilde{\delta} = \delta, \quad \tilde{\nu} = \nu + n,$$

and $\bar{X}$ is the sample mean.

Proof. It can be easily obtained by using Bayes's Theorem and arranging parameters.

In actuarial statistics, the quantity $E(X|\theta) = \theta$ is known as net risk premium (see Klugman *et al.*, 2008; among others), while expression given in (3) is the collective net premium. Now, by using Theorem (2.1), the Bayes net premium can be obtained and it is given by

$$E(\theta|\underline{X}) = \frac{\tilde{\gamma}}{\tilde{\delta}} \frac{K_{\tilde{\nu}-1}(\tilde{\gamma}\tilde{\delta})}{K_{\tilde{\nu}}(\tilde{\gamma}\tilde{\delta})}.$$

Another important estimation method is the Expected-Maximization (EM) method. This can be implemented easily following the methodology proposed in Frangos and Karlis (2004).

3. NUMERICAL APPLICATION

A numerical data set is now used to examine the fit of the new distributions. The example refers to 392 claims in a group dental basic coverage taken from Klugman (1986). The value of the log–likelihood for all the models considered, $\mathcal{E}\ \mathcal{IG}$ (exponential–inverse Gaussian distribution), $\mathcal{E}\ \mathcal{GIG}$ and $\mathcal{E}\ \mathcal{RGIG}$ results similar and is about –2804.50. These values result better than the values obtained when the Burr and Weibull distributions are considered (see Klugman, 1986) and a little worst than the values obtained when the lognormal distribution is used. Nevertheless, some differences appear when we see to thee histogram of the complete data and the fitted pdf of the different distributions (see Figure 1). This graph gives evidence than the $\mathcal{E}\ \mathcal{RIG}$ and the $\mathcal{E}\ \mathcal{RGIG}$ seems better than the $\mathcal{E}\ \mathcal{IG}$ distribution.

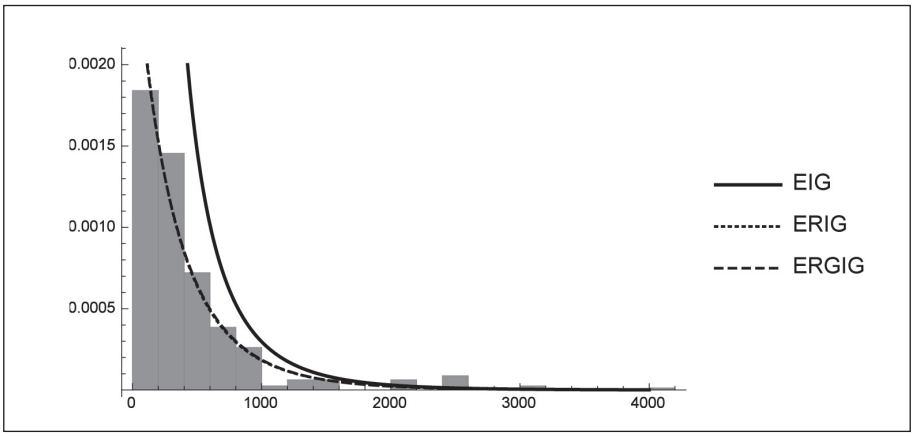


Figure 1. Distributions approximation to the histogramof the data set.

Now, the same distributions as in the previous example were considered, in order to obtain the limited expected value. The limited expected value function $L(x)$ of a claim size variable X, or of the corresponding cumulative distribution function $F(x)$, is defined by

$$L(x) = E[\min(X, x)] = \int_0^x y\, dF(y) + x(1 - F(x)), \tag{6}$$

from which the value of the function at point x is equal to the expectation of the distribution truncated at this point. Therefore, it represents the expected amount per claim retained by the insured party on a policy with a fixed amount deductible of x. The empirical limited expected values, given $\widehat{L}(x) = \frac{1}{n}\left(\sum_{x_j<x} x_j + \sum_{x_j \geq x} x\right)$, and fitted limited expected values were computed using the estimates of the parameters and plotted. The estimated values for each or the four remaining models, as well as its empirical values are shown in Figure 1. Using expression (6) we have got, after some algebra, a closed–form expression for the limited expected value when the $\mathcal{E}\ \mathcal{RIG}$ distribution is used. This is given by,

$$L(x) = \frac{1 + \gamma\delta}{\delta^2}\left[1 - \exp\left(\delta(\gamma - \phi(x, \gamma))\right)\right].$$

A similar, but more complicated expression, is obtained under the $\mathcal{E}\ \mathcal{IG}$ distribution and numerical computation is needed to obtain the limited expected values when the $\mathcal{E}\ \mathcal{RIG}$ distribution is considered.

As we can see in Figure 2 there are not differences between the fitted LEV's for the three models considered providing a good fit to the empirical values.

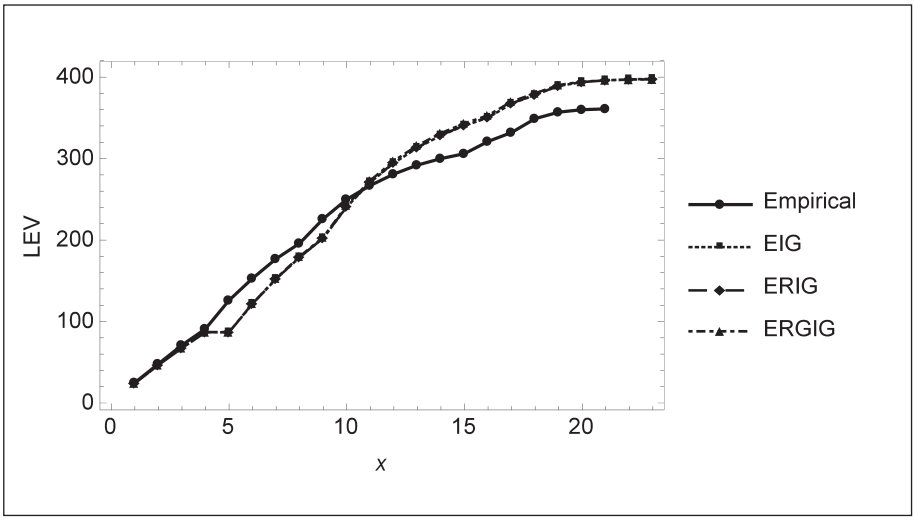


Figure 2. Empirical and fitted limited expected values for the different models considered.

4. CONCLUSIONS

This paper presents a new probability distribution well adapted to describe loss distributions: the exponential–reciprocal generalized inverse Gaussian containing the common reciprocal of the inverse Gaussian distribution as a particular case. Using its stochastic representation we obtain all common quantities of interest: density, mean, variance and survival function, among others. Computations are nowadays easy trough standard packages like

`Mathematica` and the numerical example shows that the $\mathcal{E}\ \mathcal{RGIG}$ model per forms very well in fitting loss distribution and must be incorporated in the catalogue of probability distributions to model actuarial data.

ACKNOWLEDGMENTS

Authors partially funded by grant ECO2013-47092 (Ministerio de Economía y Competitividad, Spain).

REFERENCES

Bhattacharya, S.K., and Kumar, S. (1986). "E-IG Model in Life Testing". *Calcutta Statistical Association Bulletin*, 35, 85-90.

Chhikara, R., and Folks, T. (1989). *The inverse Gaussian Distribution. Theory, Methodology, and Applications*. Marcel Dekker, New York.

Frangos, N., and Karlis, D. (2004). "Modelling Losses Using an Exponential-Inverse Gaussian Distribution". *Insurance: Mathematics and Economics*, 35, 53-67.

Klugman, S.A. (1986). "Loss Distributions". *Actuarial Mathematics. Proceedings of Symposia in Applied Mathematics. American Mathematical Society*, 31-55.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2008). *Loss Models: From Data to Decisions*. Wiley, New York.

Jørgensen, B. (1982). "Statistical Properties of the Generalized Inverse Gaussian Distribution". *Lecture Notes in Statistics*, 9. Springer-Verlag, New York.

Jørgensen, B., Seshadri, V., and Whitmore, G.A. (1991). "On the Mixture of the Inverse Gaussian Distribution with its Complementary Reciprocal". *Scandinavian Journal of Statistics*, 18, 77-89.

Seshadri, V. (1983). "The Inverse Gaussian Distribution: Some Properties and Characterizations". *The Canadian Journal of Statistics*, 11, 131-136.

Seshadri, V. (1999). *The Inverse Gaussian Distribution: A Case Study in Exponential Families*. Oxford Science Publications.

45. DETERMINATION OF THE OPTIMAL REPLACEMENT AGE FOR A PREVENTIVE MAINTENANCE PROBLEM INVOLVING A WEIBULL FAILURE PROBABILITY DISTRIBUTION FUNCTION

E. Gutiérrez-Miravete
Rensselaer at Hartford, Hartford, CT, 06120, USA,
gutieel@rpi.edu

ABSTRACT

The determination of the optimal replacement time of hardware components in mission critical assemblies is an important problem in reliability engineering. Waiting too long risks expensive liability costs due to system malfunction. Replacing items that are still in good operating condition too frequently is also costly. This paper describes the use of the symbolic manipulation program Maple for the determination of the exact solution of an optimal replacement time of a component with a Weibull-type failure probability distribution problem when using age-based preventive replacement policies. The same problem is also solved using a Monte Carlo simulation approach and the results are compared.

1. INTRODUCTION

Striving for savings and efficiency, industries often sacrifice maintenance. However, the neglect of proper maintenance of complex technological systems can involve significant risk. As an example, it has been reported that over 20% of medical device malfunctions are due to incorrect maintenance (Servomaa, 1989). Also, almost 12% of all aircraft accident reports cite maintenance as a factor and between 1994 and 2004, maintenance problems have contributed to up to 42% of fatal airline accidents in the United States alone (O'Brien, 2012). Three common types of maintenance are preventive, corrective and failure-finding. In preventive maintenance one attempts

to reduce the failure probability of the system by acting even when the item is functioning properly. Preventive maintenance can be age-based, clock-based, condition-based or opportunity based. Corrective maintenance is really repair and takes place once the item has failed and it may be impossible in mission critical systems. Failure finding maintenance is the result of failure detection during continuous monitoring. In this paper a study aimed at determining the optimal replacement age for preventive maintenance of an item whose failure probability is described by a Weibull distribution, a probability distribution function a broad general applicability. The symbolic manipulation program Maple was used to obtain an exact solution of the problem. Then a Monte Carlo simulation approach was used to obtain an approximate solution of the same problem. It is shown that both methods can be readily used to obtain solutions to optimal replacement problems encountered in preventive maintenance.

2. FORMAL DESCRIPTION OF THE PROBLEM

The failure probability distribution function of an item has been determined to be given the following Weibull-type distribution function

$$F = 1 - \exp(- (t/a)^b)$$

If the item is replaced before it fails, the cost incurred is c. On the other hand, if one waits until the item fails before replacing it, in addition to the replacement cost c one must also pay extra expenses (e.g. liability costs) k that are typically several times larger than the replacement cost (i.e. $k = r\,c$ where $r >> 1$) .

The goal is to determine the optimal replacement age t_0^* of the component under an age-based preventive maintenance policy.

3. SOLVING APPROACH

Let T be the time to failure of the component of interest. The failure probability distribution function F of the component is assumed to be well represented by a two parameter Weibull distribution, i.e.

$$F = 1 - \exp(-(t/a)^b)$$

Where a and b are the location and shape parameters of the distribution. This function is flexible and of general applicability in reliability analysis since in the limit of b=1 it becomes the exponential distribution and as b increases, it resembles a Log-Normal distribution.

The meant time between replacements of items with replacement age t_0 is given by

$$t_{MTBR}(t_0) = \int_0^{t0} t\, f(t)\, dt + t_0 \Pr(T > t_0) = = \int_0^{t0} (1 - F(t))\, dt$$

where f(t) is the failure probability density function.

The total cost per replacement period is equal to the replacement cost plus the extra cost in case failure takes place, i.e.

$$C(t_0) = (c + k\, F(t_0))/\, t_{MTBR}(t_0) = (c + k\, F(t_0))/ \int_0^{t0} (1 - F(t))\, dt$$

Note that as t_0 goes to infinity (i.e. replacement only after failure – corrective replacement) is

$$C_\infty \sim (c + k)/\, t_{MTBF}$$

where t_{MTBF} is the time between failures. If k is large then C_∞ will be large too. Note also that as t_0 goes to zero (frequent – unnecessary – replacement), the cost of replacement becomes directly proportional to the number of times the items is replaced and it can become large. Therefore, one can expect that there will be a certain item age t_0* that would yield the lowest replacement cost per replacement period (Rausand and Hoyland, 2004; Nakagawa, 2005; Osaki, 2002).

4. COMPUTATIONAL EXPERIMENTS

The software Maple was used to perform the symbolic manipulations needed to obtain the exact solution to the problem. To deal with a concrete problem the values a=10, b=3 and c=5 were selected. Using these values in maple yields the following expression for the cost per unit replacement time

$$C(t_0) = (c + k\,F(t_0))/\int_0^{t0} (1 - F(t))\,dt =$$
$$= (5 + 5\,r\,(1 - \exp(-(t_0/10)^3))/[g_1(t_0) + g_2(t_0)]$$

where

$$g_1(t_0) = 0.75\,\sqrt{10}\,t_0\,\exp(-t_0^3/2000)\,W_M(1/6,2/3,t_0^3/1000)/\sqrt{t_0}$$

and

$$g_2(t_0) = 1000\,\sqrt{10}\,\exp(-t_0^3/2000)\,W_M(7/6,2/3,t_0^3/1000)/(t_0^2\,\sqrt{t_0})$$

where WM(μ,ν,z) is the Whittaker M function of arguments μ, ν, z.

The cost per unit replacement time of corrective replacement is given as

$$C_\infty \sim (c + k)/t_{MTBF} = 0.559923\,(1 + r)$$

Figure 1 shows the values of the maintenance policy cost per unit replacement time C as a function of the component age t_0 and three

values of the multiplier factor r representing the extra cost of replacing after failure. The curves show clearly the rapid rise in cost obtained with frequent replacement of young components. The curves also show that the costs also rise with replacement age but then settle to the asymptotic maximum value of C_∞ in each case. All curves show that there is a certain component age $t_0^*(r)$ that yields the lowest cost. Further, as expected, the optimal replacement age shifts to lower values as the liability costs increase. For instance, for r=5, the optimal replacement age is $t_0^* = 4.65$ and the associated maintenance cost is C = 1.6293.

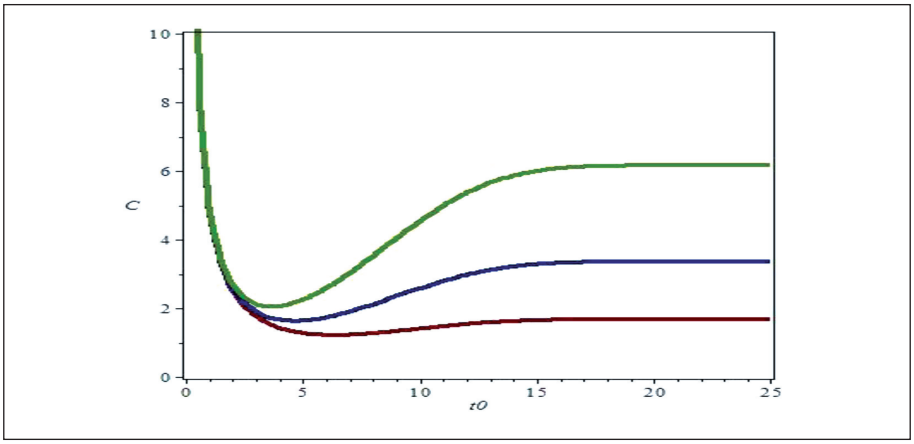


Figure 1. Computed cost per unit replacement time C versus t0 for values of r = 2, 5 and 10.

The same problem was solved using computer simulation with the Monte Carlo method. One thousand Weibull distributed random variate values were generated using the intrinsic RAND() function in the software Excel and the inverse transform method (Banks *et al.*, 2001). The computed value of the cost for t_0=4.65 is C = 1.6064 which is in reasonably good agreement with the exact value despite the small number of MC trials.

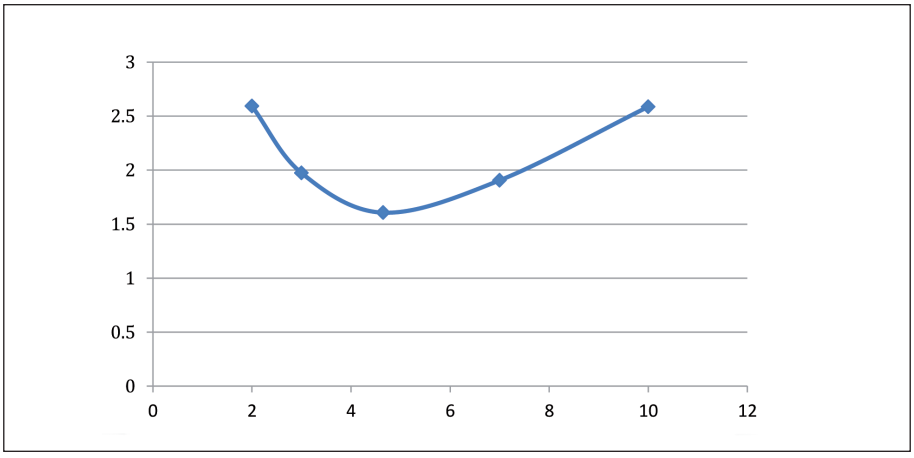


Figure 2. Computed cost per unit replacement time C versus t0 for r = 5 using the Monte Carlo method.

5. CONCLUSION

This paper demonstrates that solutions to optimal problems of age-based preventive maintenance where the failure probability distribution function is of Weibull type can be obtained by computer aided symbolic manipulation tools (i.e. Maple) and also by computer simulation using the Monte Carlo method. The results of both approaches are in good agreement and this suggests that the methods can be used with confidence to solve these types of problems.

REFERENCES

Banks J. *et al.* (2001). *Discrete Event Simulation*, Prentice-Hall, Upper Saddle River.

Nakagawa, T. (2005). *Maintenance Theory of Reliability*, Springer, London.

O'Brien, J. (2012). "When Poor Aircraft Maintenance Costs Lives". *MA-CMMS* (http://www.maintenanceassistant.com/blog/poor-maintenance- cost-lives/), accessed on February 28th, 2015).

Osaki, S. (ed.) (2002). *Stochastic Models in reliability and Maintenance*, Springer, Berlin.

Rausand, M., and Hoyland, A. (2004). *System Reliability Theory: Models, Statistical Methods and Applications*, 2nd ed., Wiley, New York.

Servomaa, A. (1989). "Mechanical Malfunctions and Inadequate Maintenance of Radiological Devices". *Health Devices* 18 (1), 53-54.

46. ON THE STABILITY OF THE REFORESTATION OPTIMAL CONTROL MODEL ASSUMPTION

D. Hampel
Mendel University in Brno, Department of Statistics and Operational Analysis, Czech Republic, qqhampel@mendelu.cz

J. Janová
Mendel University in Brno, Department of Statistics and Operational Analysis, Czech Republic, janova@mendelu.cz

J. Kadlec
Mendel University in Brno, Department of Forest and Forest Products Technology, Czech Republic, jiri.kadlec@mendelu.cz

ABSTRACT

This paper deals with validating of an assumption of the reforestation optimal control model. We focus on prices of different timber types. These serve as inputs for the simulation which gives the data for estimation of prescribed functions, whose parameters are used for the optimal control problem solution. As usual, parameters of such models are assumed as constants. Presence of trend in particular price datasets (obtained from reports of Commodity Exchange Kladno, the Czech Republic) is tested. Rate of constancy violation is assessed also by testing equality of trend slopes. Rejected are both insignificance and equality of trend slopes. Using pairwise comparisons it is possible to form sets with similar trend slopes.

1. INTRODUCTION

In many of real optimization problems one must deal with uncertainty, which means we are planning decisions without knowing what their full effects will be. Concerning static optimization techniques, there are several common approaches to cope with uncertainty in data and particularly in the parameters of the problem. Apart from stochastic programming that aims at precise incorporation of uncertainty into an optimal program, let us mention approaches that are more simple, yet often effective and appropriate: There is robust optimization that addresses optimization problems where the data is uncertain and a solution of which remains feasible despite that uncertainty is sought. Similarly, probabilistic programming focus on meeting feasibility in an uncertain environment through a minimum requirement on the probability of satisfying the constraints. The uncertainty plays even more important role in dynamic optimization, where we often deal with long-run strategies and we must consider the time evolution of the parameters of the problem. Here we must deal with stochastic dynamic programming and stochastic optimal control techniques.

However, when developing the quantitative support for real decision problems, very often the deterministic optimization models are employed and the assumption is made that the uncertain factors may be neglected in the problem. In such case, the parameters of the problem are considered to be known constants. This assumption is common both in static and dynamic optimization problems, however, omitting the uncertainty in long-run dynamic optimization may significantly challenge the results of the mathematical model.

In our contribution we focus on one particular assumption made in dynamic modelling – the assumption of constant deterministic prices. The aim is to answer to what extent such assumption is acceptable in a particular problem of forest management optimization, for which control theory is applied.

2. MOTIVATION

The final goal of our work is to solve further dynamic decision problem: to find the optimal subsidy strategy for reaching the desired forest structure in the Czech

Republic within the least time. In our model, the criterion of decision making is the economic profit from forests (including the income from selling the timber and obtaining the subsidies). To address the decision problem we introduce infinite horizon optimal control problem with free terminal points, which can be schematically described as

$$\max V = \int_0^\infty \Pi(t) e^{-\rho t} \mathrm{d}t, \tag{1}$$

where

$$\Pi(t) = \sum_{i=1}^{n} R_i(x_i(t), u_i(t)) - K_i(x_i(t), u_i(t)) \tag{2}$$

under appropriate constrains. Here, n denotes the number of tree species appropriate for the given region, $x_i(t)$ are the state variables representing the area of land forested by specie i in time t, ρ is discount rate. Further we denote total area reforested at time t by specie i (in hectare per year) $u_i(t)$, the current profit $\Pi(t)$, total revenues from selling the timber and subsidies $R_i(t)$ and total costs of growing and logging at time t $K_i(t)$. Note that we do not consider afforestation in our model. References and further discussion for this topic are stated in Hampel (2014) and Janová (2015).

We assume that

$$R_i(t) = G_i(x_i(t)) + \sigma_i u_i(t), \tag{3}$$

where G_i represents a known function of revenues from logging and $\sigma_i u_i(t)$ is the subsidy from increasing the area of specie i. For the estimation of the functions $G_i(x_i)$ and $K_i(x_i, ui)$ we adapt the assumptions made in Caparrós (2013) and set

$$G_i = g_{i0} + g_{i1} x_i + \frac{1}{2} g_{i2} x_i^2, \tag{4}$$

$$K_i = k_i + a_i x_i + b_{i1} u_i + \frac{1}{2} b_{i2} u_i^2, \tag{5}$$

where the total cost function (5) is composed from the forest management costs and reforestation costs. We assume $b_{i1} = b_{i2} = 0$ for the spruce type forest, which is supposed to be reduced. It means that decreasing the area of the forest type (after logging the current stand) is costless. To proceed with the solution of optimal control problem (1) it is necessary to estimate parameters of the functions (4) and (5) based on intensive simulation.

3. MATERIAL AND METHODS

Data about prices of different timber assortments during the years 2010-2014 were obtained from reports of Commodity Exchange Kladno, the Czech Republic. Data reports were prepared on week basis, but some weeks have no entries (different frequency for different timber assortment). We deal with totally 16 timber assortments, which cover five tree species (spruce, pine, larch, oak and beech) in several quality classes and INCOTERMS 2010 rules (see Figure 1).

For our data we will test significance of linear trend for particular datasets. Significant trends disrupt assumption about constant deterministic prices. But when the trends will have the same slope, reforesting strategy can be meaningful anyway. For this purpose we can form regression model covering all simple regressions and use general linear hypothesis H0 (see for example Milliken, 1970):

$$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{a}, \tag{6}$$

where A is an arbitrary real matrix of the type $m \times k$ and rank $m \leq k$. Further, a is a vector of the type $m \times 1$ for which equation (6) has a solution. Testing statistics for discussed hypothesis is of the form

$$F = \frac{1}{ms^2}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{a})'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{a}) \tag{7}$$

and under true H_0 it is governed by Fisher-Snedecor distribution with degrees of freedom m and $n - k$. H0 is rejected when $F > F_a(m, n - k)$. For our purposes we set $a = 0$ and

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & \cdots & 0 & 0 & 0 \\ & & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & -1 \end{bmatrix}.$$

After rejection of the global hypothesis we can use test (7) pairwise to detect differing slope parameters. It is necessary to modify significance level; we will use Šidák correction introduced in Šidák (1967). Corrected level of significance is given as

$$\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{c}}, \tag{8}$$

where c is number of compared pairs. In this context, level of significance is denoted as α for single test and α_{SID} for family of tests.

i	$\hat{\beta}_i$	L_i	U_i	R_i^2
1	2.449	2.206	2.693	0.630
2	2.573	2.158	2.988	0.581
3	4.313	3.319	5.307	0.786
4	1.764	1.107	2.421	0.321
5	2.097	1.756	2.439	0.461
6	–0.143	–1.125	0.840	0.003
7	2.862	2.277	3.447	0.443
8	4.432	3.075	5.788	0.565
9	0.257	–0.112	0.627	0.017
10	0.763	0.454	1.071	0.155
11	0.171	–0.257	0.598	0.007
12	0.494	0.261	0.727	0.078
13	–1.488	–1.952	–1.024	0.519
14	0.251	–0.030	0.532	0.020
15	0.728	0.528	0.929	0.254
16	0.812	0.616	1.008	0.319

Table 1. Estimated trend slopes $\hat{\beta}_i$ for particular timber assortments i. Tabulated are also lower (L_i) and upper (U_i) bounds of confidence interval and coefficient of determination R^2. Values of $\hat{\beta}_i$, L_i and U_i are in 10^{-3}.

4. RESULTS

Trend lines were estimated for all timber assortments. Results are illustrated in Figure 1. Significance of particular trend slopes are summarized in Table 1. We can see only 4 cases of insignificant trend slopes. Moreover, it is connected to very low value of coefficient of determination, which points to the possibility that regression line is not sufficient for trend modeling in these cases.

Based on testing statistics (7) which realizes as F = 34.686 we reject hypothesis about equality of all trend slopes (F is greater than Fisher-Snedecor quantile 1.7129). By pairwise comparisons we can detects differing slopes. We will compare c = 126 pairs, so corrected significance level is $\alpha_{SI\,D}$ = 0.000407. Based on pairwise comparisons we can establish several groups of timber assortments. The first group consists of timber with quality IV. and V. excluding pine of V. quality FCA (the reason is missing tradings in the years 2013 and 2014; EXW shipping is preferred for pine). The second group covers spruce III. B 2b EXW, spruce III. B 3 EXW, spruce III. C EXW and pine III. B 2b EXW, i.e. spruce and pine of higher quality. The third group spruce III. B 3 EXW, spruce III. C EXW, larch III. B 2b EXW and oak III. B 3 EXW (note that spruce III. B 3 EXW is also in the second group).

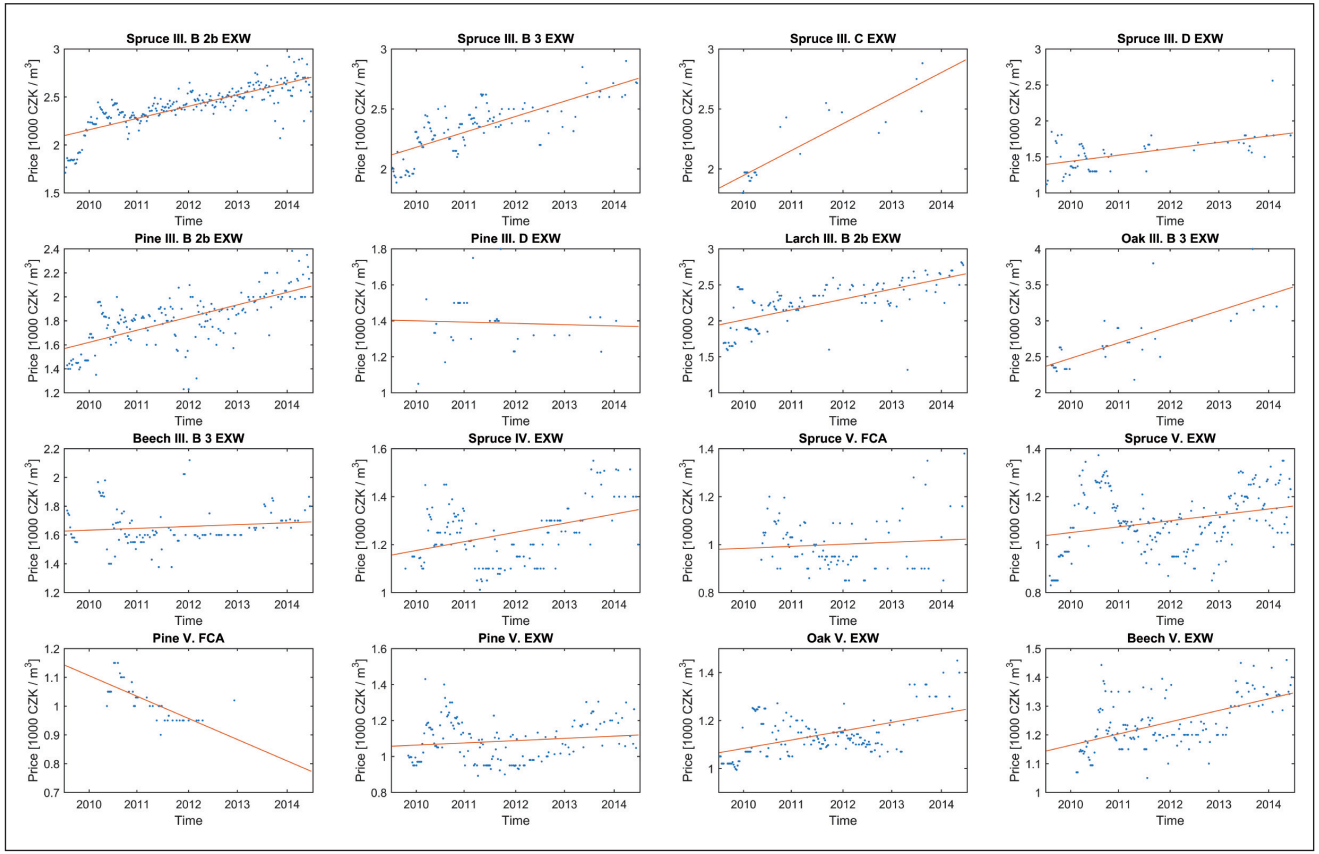


Figure 1. Development of prices of different timber assortments (data are represented by dots and estimated trends by lines).

5. CONCLUSION

Assumptions about constant prices or prices growing in the same directions seems to be false. This fact should be incorporated to our real optimal control problem in some way; it is topic for the further research.

ACKNOWLEDGMENTS

Supported by the grant No. 13-25897S of the Czech Science Foundation.

REFERENCES

Caparrós, A., Cerdá, E., Ovando, P., and Campos, P. (2013). "Carbon Sequestration with Reforestations and Biodiversity-Scenic Values". *Environmental Resource Economics*, 45, 49-72.

Hampel, D., and Janová, J. (2014). "Simulation of Data for Reforestation System". *Proceedings of the 32nd International Conference Mathematical Methods in Economics 2014*, 251-256.

Janová, J., and Hampel, D. (2015). "Optimal Managing of Forest Structure Using Data Simulated Optimal Control". *Central European Journal of Op- erations Research*, in print (DOI: 10.1007/s10100-015-0383-x).

Milliken, G.A., and Graybill, F.A. (1970). "Extensions of the General Linear Hypothesis Model". *Journal of the American Statistical Association*, 65, 797-807.

Šidák, Z.K. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association* 62, 626-633.

47. OSCILLATORY MODES ORGANIZATION IN COMPLEX TEMPORAL BEHAVIOURS

R. Herrero
Universitat Politècnica de Catalunya, Department of Physics and Nuclear Engineering, Spain,
ramon.herrero@upc.edu

G. Orriols and F. Pi
Universitat Autònoma de Barcelona, Department of Physics, Spain,
gaspar.orriols@uab.cat

ABSTRACT

Nonlinear dynamical systems can show complex temporal behaviors constituted by different characteristic oscillatory modes and frequencies. These complex temporal behaviors are developed through a mechanism based on the intertwinement of oscillatory motions emerged from Hopf bifurcations of one or more fixed points. The nonlinear mixing of these natural oscillations modes is mainly structured by the invariant manifolds connecting the different limit sets. The mixing becomes apparent in time evolutions of an attractive limit set in the form of intermittent bursts of the different oscillation modes, combined ones over the others according to the scale of frequencies. This scenario is numerically verified in systems with only one nonlinear function and an increasing number of variables (Rius, 2000). Its generalization to richer nonlinear vector fields stablishes a generic dynamical scenario with extraordinary oscillatory possibilities and can be used to tentatively explain complex oscillatory behaviors observed in nature as those of turbulent fluids and living brains (Herrero, 2012).

1. INTRODUCTION

The paper wants to shed light on the general structure of complex dynamics shown by nonlinear systems. L. Landau and E. Lifshitz (Landau, 1959) proposed the classical theory for the route to turbulence based on the linear combination as a successive incorporation of oscillatory modes associated with the dynamical activity of additional degrees of freedom.

The intuitively appealing idea that arbitrarily complex time evolutions can be achieved by combining an increasing number of simple oscillators does not correspond to any known mechanism of nonlinear dynamics, in whose mainstream the exclusive way for oscillatory mixing is widely associated with the creation of invariant tori, but the well-known theorems by Ruelle, Takens and Newhouse (Newhouse, 1978) prove the torus fragility under perturbations and the probable creation of a strange attractor. Thus, invariant tori of arbitrary order are possible but in extremely narrow regions of the space of dynamical systems only. Under this premise, the notion of oscillatory complexity should be necessarily associated with the irregularity of chaotic evolutions and nowadays the relationship between chaotic motion and complex dynamics is widely accepted.

We recover this initial idea of Landau to explain complex dynamics and turbulence as a successive incorporation of oscillatory modes. The scenario proposed in this paper differs with the Landau scenario in the unnecessary consideration of additional dimensions and the nonlinear combination of oscillation modes. In the initial proposal of Landau the number of observed oscillation modes would be half of the system dimension. Although the dynamical systems behavior becomes more and more complex as the number of involved variables is increased, the number of oscillation modes is also increased by increasing the nonlinearity of the system.

The general dynamical scenario that we call as the Generalized Landau Scenario, emerges through successive Hopf bifurcations of various fixed points that can mix with each other through the intertwinement of invariant manifolds of the corresponding limit cycles. The attractor does not need to be very high dimensional and describe complex but strictly ordered sequences of intermittent oscillatory

patterns. Under optimum circumstances, all possible oscillatory modes of the dynamical system can appear intermittently together on the time evolution. This particular behavior corresponds to the maximal or full instability of the system with the stable attractor combining all possible oscillation modes.

2. N-DIMENSIONAL SYSTEMS WITH A M-DIMENSIONAL NONLINEARITIES

We consider *N*-dimensional systems whose nonlinear part spreads in a m- dimensional subspace. The general expression can be written as

$$\frac{d\vec{x}}{dt} = A\vec{x} + \sum_{j=1}^{m} \vec{b}_j f_j(\vec{x}, \mu) \, , \, \vec{x} \in \Re^N$$

where $\vec{x}$ is the vector of variables, A is the linear matrix combining the different variables, f_j are the nonlinear functions with parameters m and $\vec{b}_j$ correspond to m linearly independent directional vectors. The nonlinear part of the vector field determines the topological structure of the potential sets of fixed points in the phase space. In general, stationary points can appear by differently oriented saddle–node bifurcations and, more rarely, through pitchfork bifurcations. They create an array of stationary points in a m-dimensional subspace, and the stable nodes have the basin of attraction defined by the stable manifolds of the surrounding saddle points. Assuming full arrays, the stable basin is surrounded by a set of 3m – 1 saddle points on the separatrix (Figure 1).

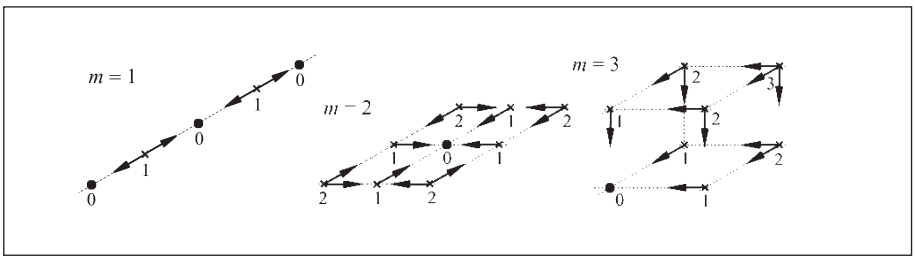


Figure 1. Schematic arrays of fixed points within m-dimensional linear subspaces. The unstable manifolds of the stationary points are represented by arrows while numeric labels denote their dimension.

The skeleton of fixed points is the origin of the different oscillation modes incorporated to the system when scanning a parameter. Consider a cluster formed by one attractive node and the full set of saddle points located on the separatrix. The first Hopf bifurcation of the stable node produces a stable limit cycle, while the succeeding bifurcations of this fixed point and the bifurcations of the saddle points will produce saddle limit cycles of different types, but all of them with a branch of their unstable manifold ending toward the attracting cycle. The number of oscillations modes created around a stationary point by Hopf bifurcations is limited by the dimensionality of its stable manifold and thus will never exceed $\frac{N-q}{2}$ where q is the dimensionality of the corresponding unstable manifold.

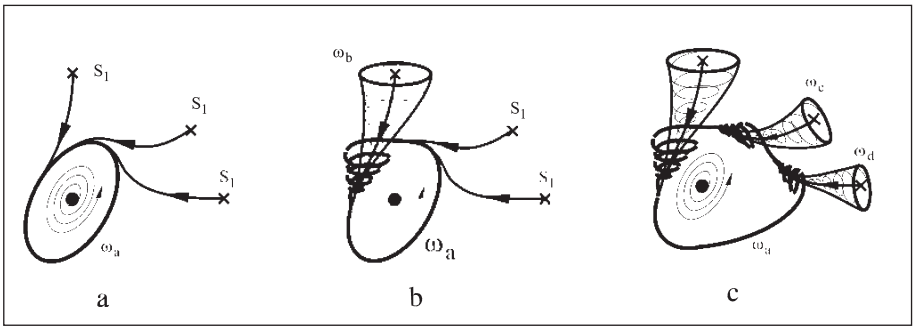


Figure 2. Nonlinear mode mixing in a set of S_1 points saddle-node connected to a given S_0 point.

The mixing process is defined in subsequent steps as follows. The first Hopf bifurcation of S_0 originates a stable limit cycle and the endings of the one-dimensional (1D) unstable manifolds of the surrounding S_1 points are transferred (Figure1a). On the other hand, S_0 has incorporated a 2D unstable manifold spiraling toward the stable cycle that will yield, in the next Hopf bifurcation, a 3D unstable manifold of the second limit cycle through which its oscillation mode will be transmitted to the stable cycle and so on for the successive bifurcations of S_0.

The first Hopf bifurcation of a given S_1 produces a saddle cycle with a 2D cone-shaped unstable manifold that ends with asymptotic tangency on the stable cycle (Figure 2b). Additional Hopf bifurcations and limit cycles can arise from S_1 point and their frequencies will be incorporated to the attractor through the con-shaped unstable manifold. The different S_1 points connected to S_0 can originate more oscillation

modes also affecting the same attractor (Figure 2c). Although the proposed scenario is not general, the main conclusions can be applied to a large number of dynamical systems.

3. VECTOR FIELDS WITH ONE-DIMENSIONAL NONLINEAR PART

We start considering systems with only one-dimensional nonlinearity, i.e. the nonlinearities are restricted in a one-dimensional subspace. This particular case corresponds to many well-known nonlinear dynamical models like the Duffins, Van der Pol, Lotka-Volterra or Rössler model.

In this case, the stationary points appear located in a straight line and the stable attractor is only influenced by the oscillation modes of a Saddle- Node pair of stationary points. For a N-dimensional system this S_0-S_1 pair can sustain up to N-1 different Hopf bifurcations which oscillations can be combined in the observed attractors.

The expected dynamics are observed in the response of a simple and passive optical system called BOITAL devices. These systems are based on the so-called opto-thermal bistability with localized absorption and have been described in detail in (Herrero, 2012; Rius, 2000) and references there in. The device basically consists of a Perot-Fabry cavity in which the input mirror is partially absorbing and the spacer is a multilayer of transparent materials with alternatively opposite thermo-optic effects. The cavity is illuminated with an input focalized laser beam and the temporal evolutions of the reflected light are analyzed. The number of different layers composing the spacer directly gives the effective dimensionality of the dynamical response.

Figure 3 presents the time evolution of the light reflected by a BOITAL device with 6 layers in the spacer showing full instability in the temporal response with 5 involved frequencies. The input power is scanned from small to large input power values, observing and increasing number of frequencies in the response. The numerical simulation from a physical model based on the thermodynamic response and opto-thermal properties of these devices show similar results. The temporal responses experimentally observed or numerically simulated are robust and can present the same structure independently of their periodic or chaotic behavior.

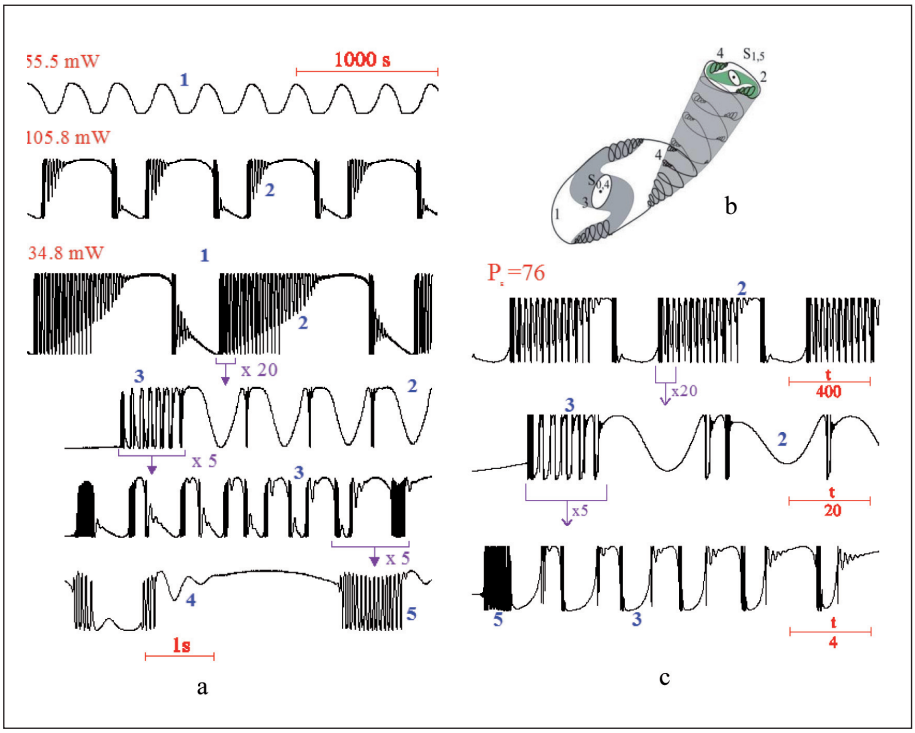


Figure 3. a) Temporal evolutions of the reflected light from a BOITAL device with a spacer composed by 6 layers at different input light powers. b) Scheme of the observed dynamics. c) Numerical simulations of the same device.

5. CONCLUSION

We proposed a mixing mechanism to explain the combination of oscillatory modes in nonlinear dynamical systems that can yield complex time evolutions in a way that evokes the Landau proposal for tentatively explaining the origin of turbulence. This behavior has been found in *N*- dimensional systems able to exploit the instability capabilities of their fixed points in full. The widely accepted idea about the necessary chaotic motion to obtain complex behavior is opposed or at least questioned by the described dynamics. The observed scenario is explained considering a general model that can include a large number of dynamical systems.

REFERENCES

Herrero, R., Pi, F., Rius, J., and Orriols, G., (2012). "About the Oscillatory Possibilities of the Dynamical Systems". *Physica* D 24, 1358-1391.

Landau, L.D., and Lifshitz, E.M. (1959). *Fluid Mechanics*, Pergamon.

Newhouse, S.E., Ruelle, D., and Takens, F. (1978). "Occurrence of Strange Axiom a Attractors near Quasiperiodic Flow on T^m, $m \geq 3$". *Commun. Math. Phys.*, 64, 35-40.

Rius, J., Figueras, M., Herrero, R., Pi, F., Farjas, J., and Orriols, G. (2000). "Full Instability Behavior of N-Dimensional Dynamical System with a One- Directional Nonlinear Vector Field". *Phys. Rev.* E 62, 333-348.

Rius, J., Figueras, M., Herrero, R., Farjas, J., Pi, F., and Orriols, G. (2000). "N-Dimensional Dynamical Systems Exploiting Instabilities in Full". *Chaos* 10, 760-770.

48. FIGHTING THE EBOLA VIRUS – AN EXERCISE IN RISK ANALYSIS AND DECISION MAKING

J. Houston
Glasgow Caledonian University, Department of Law, Accounting, Economics & Risk, Glasgow, Scotland, G4 0BA
j.houston@gcu.ac.uk.

M. Acharyya
Glasgow Caledonian University, London Campus, 40 Fashion Street, Spitalfields, London, England, E1 6PX
Madhusudan.Acharyya@gcu.ac.uk

ABSTRACT

Ebola is a deadly epidemic and has already cost the lives of at least 10,000 people in West Africa. A vaccine has yet to be discovered and the focus remains on containing the spread of the virus. Although its spread has come under control in recent months, the actions of international agencies has been criticized. In this study, we present a choice-problem, associated with the handling of Ebola by Health Policy-makers. We combine actual demographic data with realistic cost data for a limited range of interventions intended to slow transmission of the disease, into a quantitative model. The model facilitates trials of many combinations of interventions, in an effort to minimise the further spread of Ebola. The model can be used both as a teaching tool, as well as a prototype decision-making tool for the agencies involved in fighting such outbreaks, with limited budgets.

1. INTRODUCTION

The Ebola virus was first identified in the Sudan in 1976, and has remanifested itself 16 other times between then and 2012, causing 1,579 deaths and a 67% fatality rate (WHO, 2014). In 2014, the World Health Organization (WHO) declared the epidemic to be a "public health emergency of international concern" (WHO Ebola Response Team, 2014). Ebola is categorised as high-risk epidemic and the current outbreak is believed to have led to the deaths of at least 10,000 people by February 2015, mainly in Guinea, Liberia and Sierra Leone (Centers for Disease Control and Prevention, 2015; WHO, 2015; UN 2015a and 2015b). The UN has initiated a five point strategy, under the title "Global Ebola Response" (UN, 2015a, *op. cit.*): (1) stop the outbreak from spreading further; (2) treat the infected; (3) ensure essential services; (4) preserve stability; and; (5) prevent further outbreaks. The International Federation of Red Cross and Red Crescent Societies and UNICEF have also joined with the WHO to lead the implementation of the lines of actions associated with this five point strategy. However, these agencies' abilities to act are constrained by tight budgets. The UN estimates that the total effort from October 2014 to June 2015 will cost $2.27billion, of which only $1.4billion has been either spent, or pledged. Post-outbreak reconstruction will increase the likely total cost to over $4billion (UN, 2015b, *op. cit.*). The problem is exacerbated by the remoteness of the affected regions and quality of the infrastructure, that combine to affect adversely, the strategic and operational decisions facing the Agencies. Notwithstanding this, they have to justify the costs and benefits of their actions associated with the handling of this epidemic. This paper demonstrates the complexity associated with analysing risk and decision making. This is achieved by designing and presenting a risk model that exhibits an example of constrained choice that the policymakers face.

2. THE RISK MODEL

In this context, we refer to *risk* as the risk of the virus continuing to spread both within each region, and into other, previously unaffected regions. We score the risk of Ebola being transmitted to other victims in the regions of five West African countries: Guinea, Sierra Leone, Liberia, Mali and the Ivory Coast, according to pre-selected risk-attracting attributes. Some of these pertain to the whole country (e.g. GDP, Public

Expenditures, Literacy/Death Rates and number of Ebola Treatments Centres, ETCs). Others pertain to the region itself (location to Capital City, population densities), according to data availability. Each component yields a score, emanating from its own risk table and the 'value' possessed by the region at that point in time. The relative importance of each component is expressed through the absolute values of the scores. For example, Table 1 summarises the scores awarded for location to the nation's capital city. The sum of these attribute scores gives a total risk score, which is then mapped to a specific qualitative risk class. For illustrative purposes, it is assumed that each region will exhibit one of *four* levels of risk, as indicated in Table 2.

This gives the 'opening' risk position for each region. It is assumed that the Agencies have at their disposal, *three* risk-reducing interventions that can be offered to each region, within the confines of the available budget.

Location	Risk Score
Adjacent	10
Close	50
Distant	100

Table 1. Ebola Risk: Location to Capital City Risk component.

	Risk Category	Threshold Risk Score
1.	Moderately Serious	0
2.	Serious	400
3.	Very Serious	750
4.	Extremely Serious	900

Table 2. Ebola Problem: Qualitative Risk Classes.

From previous experience, these are believed to reduce risk *scores* with the possibility that their implementation may be able to move a region into a lower risk *class*. These are: (1) New ETCs, each staffed with care workers and equipment, including a laboratory; (2) Burial and Sanitization teams to safely dispose of victims' remains and to disinfect their living spaces; (3) Public Education Programmes (PEPs), to educate citizens in how to detect the early signs of Ebola and take precautions to reduce the risk of acquiring, or transmitting it. In this model, a region can be offered an ETC, whose effectiveness is determined by the number already there (Table 3).

# ETCs already in the Region	% reduction	# ETCs already in the Region	% reduction
0	25%	4	2%
1	10%	5	1%
2	8%		
3	5%	6	0.5%

Table 3. Impact on the Risk Score of the construction of another ETC.

A Burial and Sanitization Team's effectiveness is determined by the physical territory (in km^2) it has to cover. This territory is a function of the region's overall size and the number of teams to be deployed there. Accordingly, the model calculates the average territory implied as it considers the number of teams to deploy, and thence the region's component score, according to Table 4.

Area (km^2) per team	% reduction in score	Area (km^2) per team	% reduction in score
10-99	25%	10000-99999	3%
100-99	15%	100000-plus 0%	0%
1000-9999	8%		

Table 4. Impact on the Risk Score of the Burial and Sanitisation Team density.

The Agencies responsible for Public Health have to decide on whether to offer any form of public health awareness programme in each region. If they do decide to, then there are two levels of programme available: 'Low' and 'High', with different per-capita costs and effectiveness in reducing the region's total risk score. 'Low' reduces the region's score by 15%, 'High' by 25%. Of course, these Interventions are not costless, and the budget available to the Agencies is, perforce, limited. Different costs were assumed for each country, to take into account the different terrains, labour markets, etc in each one. The absolute costs of the ETCs dwarf all the other interventions, though it should be borne in mind that they are either *per-team* or *per-capita* costs. Nevertheless, the per-dollar effectiveness of each Intervention is designed, deliberately, to be very different in order to provoke a debate to the extent to which 'expensive' interventions can be ignored as part of the solution. We use Palisade's Risk Optimizer program and Excel, to trial and cost, combinations of ETCs, Burial-Sanitization Teams and PEPs in each region. A budget of $4billion was assumed, being broadly in line with the UN's current estimate as to the total cost of

fighting Ebola. However, this is deliberately less than the total cost of giving each region an additional ETC, a large number of Teams and High-level PEPs, otherwise, there would be no decision problem. When trialling each combination of interventions, the model computes the new risk score and corresponding risk class for each region. If the percentage reduction is sufficient, a region will be moved into a lower Risk Class (Table 2). The objective is to maximise the total number of changes across all 64 regions, subject to not exceeding the $4billion budget. In order to reduce the search for the optimal solution, it was assumed that a region can: (1) receive, at most, one new ETC; (2) receive a maximum of 50 Burial & Sanitization teams, regardless of its physical area and/or population; (3) be considered for either a Low- or High-level PEP.

3. RESULTS

The model was then run for 18 hours, on a Dell Optiplex 7020 PC, in which time it performed 13.21 million combinations of interventions[1]. The best combination found in the time allowed results in a total of 76 Risk-class reductions, affecting 55 of the 64 regions. Of those 55, two regions were moved down by three risk classes (the maximum possible); 17 by two and 36 by one. The average cost of each risk class improvement is $54.5 million, though with a large standard deviation ($67.7m). Table 5 summarises the number and costs of ETCs, Burial and Sanitization Teams and PEPs to be deployed into the regions.

Interventions	#	Cost ($m)	People reached (m)	Spend per person	% of total
ETCs to be built	23	$3,679	18.679	$196.96	96.6%
Burial & Sanitization teams	2,189	$1.58	48.444	$0.03	0.0%
PEP	24 (High) 27 (Low)	$129.59	36.337	$3.57	3.4%
Total Spent $3,810.17m		Total Unspent $189.83m			

Table 5. Summary Results: all areas.

[1] The spreadsheet showing the detailed computation and simulation are available from the Corresponding Author, upon request.

Table 6 summarises the distribution of regions into the four risk classes before and after the Interventions. Remembering that the assumed costs were not actually 'real', but 'realistic', we can see that the expensive ETCs account for almost all the $3.8billion spent. The apparent cost per person reached by the 23 ETCs is also considerably greater than both the Burial Teams and PEPs. This may cast some doubt on the wisdom of devoting so much of the budget on the ETCs, a point that is used to provoke debate amongst the students. Is it acceptable to spend so much money on high-profile interventions, such as ETCs when lower-cost solutions, might be more 'dollar-effective'?

Risk Categories	Before	After
Moderately Serious	0	25
Serious	32	28
Very Serious	2	11
Extremely Serious	30	0

Table 6. Risk Categories before- and after Interventions.

In assessing how good the Optimizer is in finding the optimal solution, we note that even after 18 hours run-time, $190million remains unspent. We might consider purchasing another ETC in Guinea, Mali or Sierra Leone, or thousands more Burial-Follow-up teams, and/or upgraded PEPs. Any seven of the nine regions left in their original classes (all 'Serious' in this case), could be moved down to the 'Moderately Serious' category if an extra ETC were committed. However, these are all in countries where an ETC costs more than the unspent budget. Manually adjusting the Burial Teams and PEPs does, of course, use up some of the unspent budget, though it was not possible to obtain any additional risk class changes. However, the Optimizer does suggest committing over $9million on five of the nine 'no-change' regions. This money could be added back into the unspent portion and the resulting $199million used to try and find further improvements. It may be that the Optimizer just had not got around to doing that when the time permitted to find a solution, had elapsed.

4. CONCLUSION

It is claimed that the probabilities of spreading Ebola across other parts of the world and the size of the associated impact are totally unknown and unpredictable. Moreover, no effective cure has yet been discovered to treat the Ebola infected patients successfully. Consequently, the mitigation of risks associated with spreading Ebola is the only way to manage this emerging risk. In this paper, we have offered a practical approach to the choice problem associated with the handling of the Ebola virus by health policy makers. This practical exercise of mitigating the risk of spreading Ebola may contribute to the policy development and implications' issues both a public and private levels (e.g. Health Insurers). In addition, the exercise can be used in teaching resource- constrained optimisation approaches to Risk Students.

REFERENCES

Centers for Disease Control and Prevention (accessed 10th February, 2015). http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/ case-counts.html

United Nations (2015a). http://ebolaresponse.un.org/ebola-response, accessed up to 10th February, 2015.

United Nations (2015b). Ebola Outbreak: *Updated Overview of Needs and Requirements for January-June 2015*.

WHO Ebola Response Team (2014). "Ebola Virus Disease in West Africa – The first nine Months of the Epidemic and Forward Projections". *New England Journal of Medicine*, 16, 371.

World Health Organisation (2014, 2015). Ebola Situation Reports, (http://apps.who.int/ebola/en/current-situation/ebola-situation-report), accessed 10th February, 2015.

49. SURVIVAL AND QUALITY OF LIFE

C. Huber-Carol
University of Paris Descartes, Mathematics Department, France,
catherine.huber@parisdescartes.fr

ABSTRACT

Survival data analysis and reliability have many features in common. But the available data have often different properties. The patients are not only "objects" of study but also "subjects" who may decide at any time to stop their participation in the experiment, and for whom we cannot accelerate the onset of the expected event for the sake of avoiding right censoring by the date of the end of the study. Moreover, truncation is added to censoring (Huber-Carol *et al.*, 2009) as sometimes only a part of the targeted sample is observed. Cox model (Cox & Oaks, 1998) is the favorite model to analyze biomedical survival data. Most of the incriminated properties of this model may be relaxed, like for example, the assumption that the impact of a factor is constant in time, leading to extended Cox models (Bagdonavicius & Nikulin, 2002; Cox & Oaks, 1998; Huber-Carol & Nikulin, 2008).

The aim of survival analysis is not only to increase the length of the pure survival time (to death) but also to improve its quality. For that purpose, we consider processes leading to different possible health states (Huber-Carol & Pons, 2005) of a patient, modeled using semi-Markov processes, and also we estimate the impact of pollution on the number of "years free of disease" lost, for the sake of prevention, using First Hitting Time models (Lee & Whitmore, 2006, 2010; Chambaz *et al.*, 2014).

1. A SINGLE EXPECTED EVENT

X is the random time elapsed before an event of interest occurs, $S(t) = P(X \geq t)$ its survival function, $f(t)$ the corresponding density, assumed to exist, and $h(t) = f(t)/S(t)$ its hazard rate. Very often, X is not observed but it is only known to be bigger than some value C, and C is called the censoring variable. This occurs for example if the expected event is death from some disease and the patient is still alive when the study stops. This is called right censoring. But some other censoring schemes may happen, like left censoring or interval censoring. Moreover, some patients may totally escape the sample and this is called truncation.

Censoring and truncation

Definition and examples of censored and/or truncated data in survival analysis are given. The most usual case treated in the literature is right censoring, but many other possibilities may occur so that we were lead to study the most general case of censoring and truncation (Huber-Carol *et al.*, 2009). X is said to be truncated by a set B and censored by a set $A \subset B$ if X is observed only conditionally on being in B and X is known to be in $A \subset B$. Observation is $(A, B), A \subset B$, and the likelihood is proportional to $P(A)/P(B)$. Based on AIDS acquired by transfusion, an example is provided of left censoring together with right truncation.

Cox model and its extensions

The model mostly used in medical applications is Cox model which defines the hazard rate $h(t)$ as a special function of the available covariates $Z = (Z_1, \ldots, Z_p)$

$$h(t|Z) = h_0(t)exp(< \beta, Z >) \tag{1}$$

where h_0 is the baseline hazard, obtained for $Z = (0, \cdots, 0)$, Z is a vector of covariates in R^p which have an impact on the survival time, like treatment, biological measurements, and environmental features. The success of this model is due to the fact that it allows through the partial Cox likelihood (Cox & Oaks, 1998), to estimate easily the parameter β, without taking care of the baseline hazard h_0 considered as a nuisance infinite dimensional parameter, and using all observations, even the censored ones. This very attractive model may be refined on several respects: the linearly exponential

effect of Z may be changed to some function $q(Z)$; the chronological time, which can have an influence as therapies may be improved as time passes by, may be included in the model so that two different times are taken simultaneously into account; also, the fact that the effect of each covariate is supposed to be constant in time can also be discarded; moreover, by putting the integrated hazard $H(t) = \int_0^t h(u)du$ inside the model as a covariate depending on time and representing the wear of the subject, one can get a more complex model.

Accelerated Failure Time Models (AFT)

Accelerated models (Bagdonavicius & Nikulin, 2002) are more specifically used in reliability theory than in medical studies as it is not possible to "augment" the stress of patients in order to induce the expected event within a shorter time to avoid right censoring. But, still, one could estimate, by using such models, the acceleration of the onset of a disease (like lung cancer for example) due to the exposure to an environmental or behavioral factor (like smoking for example). Several accelerated failure time models are presented.

2. SEVERAL EXPECTED EVENTS

As we are interested in the quality of life of patients, and not only in their survival, there are several health states to consider and the influence of several factors (treatments, biological, environmental and behavioral data) on the transition from one state to another one and also on the time spent in each specific state.

Sequential case: multistate processes

One way to proceed to study the transitions from one state to another, that is more general than the Markov assumption, is to use semi-Markov models. In such models, the duration between the onset of two successive states depend not only on the initial state, but also on the final one. Such processes have been studied when, due to right censoring, you do not know which state would have been the final one after the censoring took place. Semi-Markov processes verify

$$P(X_k \leq t, J_k = j' | J_0, .., J_{k-1} = j, X_1, .., X_{k-1}) = P(X_k \leq t, J_k = j' | J_{k-1} = j). \quad (2)$$

They can be defined through the following two sets of parameters

1. The direct transition probabilities from state *j* to state *j'* :

$$p_{jj'} = P(J_k = j' | J_{k-1} = j), \quad j' \in J(j). \quad (3)$$

2. The law of the sojourn time in state *j* leading to *j'*:

$$F_{|jj'}(t) = P(X_k \leq t | J_{k-1} = j, J_k = j'). \quad (4)$$

For a semi-Markov model, the non parametric maximum likelihood (NPML) estimator of the cumulative hazard is derived (Huber-Carol & Pons, 2005) taking into account the right censoring that occurs for the initial and final states.

Parallel case: correlated survival data

The quality of life depends also on genetic factors, so that studies have been done regularly on twins in order to be able to separate what is due to genetic factors from what is due to the environment. The survival times thus studied, which is also true for litters if we now go from medicine to biology, are correlated and this has to be taken into account.

3. NUMBER OF YEARS FREE OF DISEASE

A frequent problem arises when people who have been submitted to an occupational exposure to pollution should receive a compensation computed through the number of years free of disease that they lost due to their exposure. So, the statistician has to separate from other causes of onset of the disease (like habits of consumption, genetical predisposition,...) the role of this occupational exposure. One way of doing this is to use a First Hitting Time (FHT) model (Lee & Whitmore, 2006, 2010) in the following way. The health state $H(t)$ of the patient at time t is a latent (not observable) variable that is a process which depends on many factors, one of which

is this professional exposure. Starting from an initial value H_0, function of initial variables (genetic features, date of birth,...), the process *H(t)* is assumed to be a Brownian process with a negative trend depending on the lifelong variables of the subject (habits of consumption, biological data...) until this amount of health *H(t)* reaches the boundary 0. In this setting, the occupational exposure is supposed to accelerate the time to reach the 0 boundary. Of course, the amount of health is relative to the particular disease under consideration. The provided example is about lung cancer and asbestos exposure, so that the 0 boundary is reached by H at the time t when lung cancer occurs, or, to be more precise, when it is diagnosed. Estimating this acceleration allows the computation of the number of years free of disease lost by the patient (Chambaz *et al.*, 2014).

The amount of health relative to lung cancer is represented as a stochastic process *H(t)*:

$$H(t) := H(t|h,\mu) = h + \mu t + B(t) \tag{5}$$

where

1. $h > 0$ is the initial amount of health
2. $\mu < 0$ is the slope of the process
3. *B(t)* is a Brownian motion

Time to onset T*(h, μ)* of the disease is defined as the first time *H(t)* hits 0:

$$T(h,\mu) = \inf\{t \geq 0 : H(t|h,\mu) \leq 0\} \tag{6}$$

T (h, μ) is $< \infty$ *as* $\mu < 0$) and its distribution is inverse Gaussian:

$$F(t|h,\mu) = 1 + e^{-2h\mu}\Phi\left((\mu t - h)t^{-1/2}\right) - \Phi\left((\mu t + h)t^{-1/2}\right),$$

where Φ is the standard normal cdf. The presence of occupational exposure to asbestos accelerates the time to onset *T* by a nondecreasing continuous function *R* on $\mathbb{R}+$ such that $R(t) \geq t \; \forall t$:

$$T(h,\mu,R) = \inf\{t \geq 0 : H(R(t)|h,\mu) \leq 0\}, \tag{7}$$

so that $T(h, \mu, R) \leq T(h, \mu)$, and

$$T(h, \mu, R) \geq t \quad \Leftrightarrow \quad T(h, \mu) \geq R(t),$$

The two parameters h and μ are functions of initial and lifelong covariates observed on the patients. Then, the number of years free of disease lost due to asbestos exposure of each patient is derived from his/her estimated acceleration of the onset of the disease due to his/her specific exposure.

REFERENCES

Bagdonavicius, V., and Nikulin, M. (2002). "Accelerated Life Models, Modeling and Statistical Analysis". *Monographs on Statistics and Applied Probability*, 94, Chapman & Hall, 344 p.

Chambaz, A., Choudat, D., Huber-Carol, C., Pairon, J.C., and Van der Laan, M.J. (2014). "Analysis of the Effect of Occupational Exposure to Asbestos Based on Threshold Regression Modeling of Case-Control Data". *Biostatistics*, 15, 2, 327-340.

Cox, D.R., and Oaks, D. (1998). "Analysis of Survival Data". *Monographs on Statistics and Applied Probability*, 21, Chapman & Hall, 201 p.

Gross, S., and Huber-Carol, C. (2008). "Multivariate Survival Data with Censoring". *Statistical Models and Methods for Biomedical and Technical Systems*. Birkhauser, 55-64.

Huber-Carol, C., Solev, V., and Vonta, F. (2009). "Interval Censored and Truncated Data: Rate of Convergence of NPMLE of the Density". *Journal of Statistical Planning and Inference*, 139, 1734-1749.

Huber-Carol, C., and Nikulin, M. (2008). "Extended Cox and Accelerated Models in Reliability with general Censoring and Truncation". *Statistical Models and Methods for Biomedical and Technical Systems*. Birkhauser, 3-21.

Huber-Carol, C., and Pons, O. (2005). "Independent Competing Risks versus a General Semi-Markov Model". *Far East Journal of Theoretical Statistics*, Pushpa Publishing House, 16 (1), 87-104.

Lee, M.L.T., and Whitmore, G.A. (2006). "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary". *Statistical Science*, 21 (4), 501-513. ISSN 0883-4237.

Lee, M.L.T., and Whitmore, G.A. (2010). "Proportional Hazards and Threshold Regression: Their Theoretical and Practical Connections". *Lifetime Data Analysis*, 16 (2), 196-214. ISSN 1380-7870. doi: 10.1007/s10985-009-9138-0.

50. POVERTY AS A RISK FACTOR FOR SUSTAINABLE DEVELOPMENT

L. Issever Grochová
Mendel University in Brno, Faculty of Business and Economics,
Department of Economics, Czech Republic,
ladislava.grochova@mendelu.cz

ABSTRACT

The paper deals with actually discussed topic of sustainable development. Its aim is to reinforce the argumentation for the inverse U shape of Environmental Kuznets Curve basing it on microeconomic indifference curve analysis. Widely used Cobb-Douglas functional form is applied to describe individual preferences over wealth and pollution. In particular, positive preferences over wealth while neutral for poorer and negative preferences for richer individuals over pollution are expected. Data on income (real GDP per capita), and pollution (CO_2 emissions) from 248 states enters the direct utility function of Cobb-Douglas form. Nonlinear least-squares estimator is used. We demonstrate that individuals in the poorest countries do not take pollution into account when maximizing their utility, while higher income renders individuals sensible to polluted environment. To conclude, our results confirm that the irrelevance of pollution in decision-making for poor individuals is a potential risk factor for sustainable development.

1. INTRODUCTION

The discussions on the link between economic performance and environment are based on the concept of sustainable development founded in the 1970s when some studies pointed out the mismatch between economic and ecological goals (see e.g. Meadows *et al.*, 1972). Since then the importance of environmental issues has grown

rapidly (Issever Grochová, 2014, 2015). The link between economic performance and environment is traditionally described with the Environmental Kuznets Curve (Grossman and Krueger, 1991, 1995) that is a non-linear function of pollution dependent on income level. Some studies (e. g. Andreoni and Levinson, 2001; Grossman and Krueger, 1993; Halkos, 2003; Mills and Waite, 2009) confirm that it follows an inverted U shape since at lower income levels the use of natural resources is elevated due to prevalent heavy industries results in more pollution, while if a certain level of income is achieved the awareness of the problems due to the environmental degradation increases stimulating demand for and consequently supply of environmental quality. The research concludes that it is sufficient to focus primarily on economic performance enhancement as economically stronger subjects are more interested in pollution abatement. This means that poverty can be regarded as a risk factor for such an economic development that is friendly to environment called sustainable development (Strelec *et al.*, 2014).

The macroeconomic goals are hardly fulfilled without their microeconomic foundation. The abovementioned so implies that with higher standard of living individuals become more environmentally sensitive and they start to protect the environment as this issue becomes relevant for their goal of utility maximization.

The objective of the paper is then to show that the state of environment is negligible for decision/making for individuals with low incomes while relevant for high-income individuals. In terms of preferences, income is always positively evaluated and as the income increases the attitude to pollution changes from neutral to negative. The aim is then to provide microeconomic argumentation for U-shaped Environmental Kuznets Curve.

The individuals can also be directed towards the desired behaviour through formal rules. The creation of strong institutions as an effective tool of pollution abatement-economic enhancement confirm e.g. Andreoni and Levinson (2001), Bhattarai and Hammig (2004), Dasgupta *et al.* (2001), Issever Grochová (2014), Leitão (2010), or Magnani (2000). However, the enforcement of the rules is often accompanied with inadequate high transaction costs. Thus, we assume that it should be efficient to target primarily at economic performance.

2. METHODOLOGY AND DATA

To illustrate individual preferences over wealth and pollution indifference curve analysis is used. Let u be the (direct) utility function which follows Cobb-Douglas functional form for a two-good situation with a negatively evaluated independent variable (Weber 2001):

$$u : \Re_+^n \to \Re \text{ with } u(x) = x_1^{\alpha_1} x_2^{\alpha_2} = \prod_{i=1}^{2} x_i^{\alpha_i} \quad (x_1 \geq 0, x_2 < 0, \alpha_1 > 0, \alpha_2 < 0)$$

where α i are parameters, x_1 and x_2 vectors of incomes and pollution, respectively. We assume that u satisfies the basic axioms of consumer theory (see Voorneveld 2008). As our hypothesis expects positive preferences over income and neutral or negative preferences over pollution we can write:

$$\frac{\partial u}{\partial x_1} = \alpha_1 x_1^{\alpha_1 - 1} x_2^{\alpha_2} = \frac{\alpha_1 u}{x_1} > 0,$$

$$\frac{\partial u}{\partial x_2} = \alpha_2 x_1^{\alpha_1} x_2^{\alpha_2 - 1} = \frac{\alpha_2 u}{x_2} = 0, \text{ for } x_1 < x_1 *$$

and

$$\frac{\partial u}{\partial x_2} = \alpha_2 x_1^{\alpha_1} x_2^{\alpha_2 - 1} = \frac{\alpha_2 u}{x_2} < 0 \text{ for } x_1 \geq x_1 *,$$

where x_1* is the income threshold from which the preferences over pollution change from neutral to negative.

Our dataset comprises 248 world economies in 2012 and includes GDP per capita (constant US$, x_1), CO_2 emissions (metric tons per capita, x_2) and Happy planet index (u) as a proxy for utility assuming that an individual is a representative economic agent. Countries are classified by income according to World Bank classification[1] (World Bank, 2012). For the estimation nonlinear least-squares estimator is used. The initial values of parameters are set up to: $\alpha_1 = 1$ and $\alpha_2 = 0$.

[1] Low income: $1,025 or less, Lower middle income: $1,026 to $4,035, Upper middle income: $4,036 to $12,475, and High income: $12,476 or more.

3. RESULTS AND DISCUSSION

The estimations of parameters α_1 and α_2 proceeded as follows. Convergence was achieved after 12, 8, 9, 10 and 12 iterations in models for all, low income, lower middle income, upper middle income and high income countries, respectively. As referred in Table 1 estimated parameters are of expected signs and are statistically significant at 1% level.

	All	Low income	Lower middle income	Upper middle income	High income
α_1	0.448 (0.002)***	0.562 (0.006)***	0.497 (0.003)***	0.468 (0.005)***	0.394 (0.007)***
α_2	-0.363 (0.011)***	-0.011 (0.002)***	-0.117 (0.021)***	-0.255 (0.034)***	-0.197 (0.032)***
N	581	131	140	144	166
R-sq	0.850	0.933	0.949	0.949	0.955
adj.R-sq	0.850	0.932	0.948	0.948	0.955
BIC	4928	981	1061	1094.1	1203.8

Note: Standard errors in parentheses, * p←0.05, ** p←0.01, *** p←0.001.

Table 1. Cobb-Douglas utility function estimations.

Alphas 2 are either close to zero or negatively signed which shows neutral or negative preferences over pollution. That is low income countries show the unimportance of environment for individual decision-making. As the income increases, pollution becomes relevant in the utility maximization process. In particular, pollution must be compensated with higher income otherwise utility decreases. Whereas positive alphas 1 confirm that the magnitude of income is an important attribute of happiness.

Also from Figure 1 it is evident that polluted environment is neutral for individuals in low income countries while negatively valued by all other income classes. For a given income poorer countries would be willing to tolerate more polluted environment.

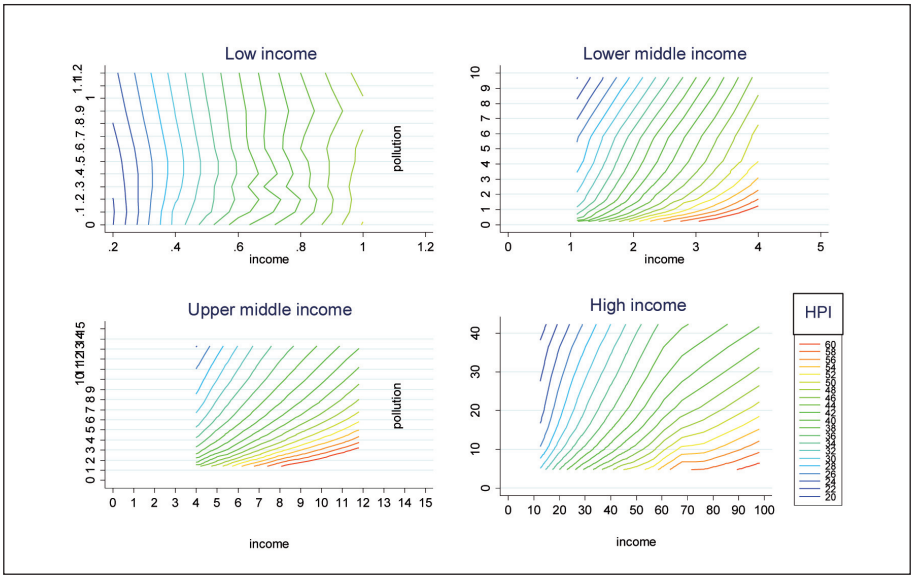


Figure 1. Indifference curves of individuals in particular countries (utility, i.e. HPI is scaled ascendingly from blue to red).

It is also interesting to compare the individuals across 4 mentioned income classes. Assuming given pollution, it is evident that individuals from poorer countries reach the same level of utility as richer representatives at lower income. Furthermore, the changes in utility caused by changes in income are smaller for individuals from richer countries (ranging from 0.394 in high income countries to 0.562 in low income countries). In terms of quantity of variables consumed, it is more and more demanding to reach higher level of utility mainly for individuals in higher income countries as the sum of α and β diminishes with income.

4. CONCLUSION

The aim of the paper was to demonstrate that preferences over environmental issues change with wealth and provide so microeconomic argumentation for U-shaped Environmental Kuznets Curve. In particular, individuals have neutral preferences over pollution if their income is $1,025 or less. From this income threshold individuals

evaluate pollution as negative. Consequently, when following their goal of utility maximization they are more interested in pollution abatement than individuals from the poorest countries who are entirely insensible to pollution. Therefore, poverty can embody potential thread as individuals from the poorest countries ignore the problems of pollution. This conclusion offers an interesting policy recommendation that is to enhance economic development and so to initiate the switch from neutral to negative preferences over pollution in individuals.

Although individuals can also be directed towards the desired behaviour through formal rules the enforcement of which is often accompanied with high transaction costs, based on our results we state that it is efficient to target first at improvements in economic performance.

Another finding is that individuals from lower income countries reach the same level of utility as those in higher income countries at lower income. Moreover, due to (positive) changes in income utility grows more rapidly in poorer countries. This confirms Easterlin paradox (Easterlin, 1974, 1995) which states that substantial real income growth has not been accompanied over the last fifty years with any corresponding rise in happiness levels in the countries analysed.

REFERENCES

Easterlin, R.A. (1974). "Does Economic Growth Improve the Human Lot? Some Empirical Evidence". In David, R., and Reder, R. (eds.), *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*. New York: Academic Press.

Easterlin, R.A. (1995). "Will Raising the Incomes of All Increase the Happiness of All?". *Journal of Economic Behavior and Organization*, 27, 35-47.

Grossman, G.M., and Krueger, A.B. (1991). "Environmental Impact of a North American Free Trade Agreement". *Working Paper 3914*. National Bureau of Economic Research, Cambridge, MA.

Grossman, G.M., and Krueger, A.B. (1995). "Economic Growth and the Environment". *Quarterly Journal of Economics*, 110, 353-377.

Issever Grochová, L. (2014). Regulatory quality and sustainable economic development. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 6.

Issever Grochová, L. (2015). "Can Economic Growth Be Acceptable if Environmental Issues are Considered? Economic Consequences in the EU Countries". In Simos, T. E., *12th International Conference of Numerical Analysis and Applied Mathematics 2014*: ICNAAM 2014, AIP Conference Proceedings. Melville, New York: American Institute of Physic. In press.

Střelec, L., Hampel, D., Issever Grochová, L., and Janová, J. (2014). "On the Measurement of Sustainable Economic Performance". In *Proceedings of the 32nd International Conference Mathematical Methods in Economics 2014*. Olomouc: Palacký University, 985-990.

Vooreneveld, M. (2008). "From Preferences to Cobb-Douglas Utility". *SSE/EFI Working Paper Series in Economics and Finance*, 701. October 27, 2008.

Weber, C.E. (2001). "A Production Function with an Inferior Input". *The Manchester School*, 69, 612-622.

World Bank (2012). *World Bank List of Economies* (July), http://www.healthsystemsglobal.org/Portals/0/files/World_bank_list_july2012.pdf.

51. MEASURING RISK IN COSTING SYSTEMS USING COSTING AT RISK (CaR)

V. Jiménez
University of Minho, Department of Production and Systems, Portugal, and University of Valle, Department of Accounting and Finance, Colombia,
victor.jimenez@dps.uminho.pt

P. Afonso
University of Minho, Department of Production and Systems, Portugal,
psafonso@dps.uminho.pt

ABSTRACT

Risk management is the process of measuring risk in order to develop and implement strategies to manage that risk. Risk analysis in costing models recognizes that data used in costing systems is uncertain. Traditionally, researchers and practitioners have been focused on deterministic costing models without acknowledge and handle conveniently such uncertainty. This paper presents and discusses a methodology for measuring the risk within costing systems: Costing at Risk (CaR). For the determination of CaR a four steps methodology is proposed. The analysis of manufacturing costs is used to explain its application. CaR permits to identify the components which most influence the variability of cost objects allowing the design of more effective strategies of cost risk mitigation.

1. INTRODUCTION

Costing models are important for firm's managers in day-to-day activities as well as for strategic purposes. Furthermore, such models can be used for cost analysis from both ex-ante and ex-post perspectives namely, for budgeting and bidding

activities and product design decisions, for production and process cost control, among others. In this context, several conceptual approaches have been developed to support costing systems, ranging from simple standard costing to more sophisticated models such as activity based costing. Indeed, there are different solutions for different situations. The appropriateness of these tools and consequently cost management effectiveness result from fitting the characteristics of the organization and the external context, taking into account several contingencies such as organizational structure, external environment, size, strategy, among others (Donaldson, 2001).

In general, costing systems are viewed as deterministic models. Nevertheless, the accuracy of the computed cost can be significantly influenced by initial estimates which themselves are uncertain and might lead to inaccurate information for decision making. Facing uncertainty and imprecise information, the usual response is to manage uncertainty using a framework. The management of uncertainty can be termed as risk management (Esmalifalak, Albin and Behzadpoor, 2014) which implies a systematic process of identification and assessment of risks.

In this paper it is purposed "Costing at Risk" (CaR) to deal with uncertainty in costing models. Process and product variability can have a considerable impact on budgeting processes and cost management effectiveness in a company. Thus, it is necessary to design tools that let manage that risk. This paper presents the methodology for the computation of CaR, making an analogy with Value at Risk (VaR), using three different methods, Historical Simulations, Delta-Normal Approach and the Monte Carlo Simulation. Managerial implications and advantages and disadvantages of the method and the potential of its application are discussed at the end of the paper.

2. COSTING AT RISK

Costing at Risk or CaR can be defined as the maximum expected cost for a product or service for a future period, given a certain level of confidence. For its determination the following steps are proposed:

1. Build the model for the determination of costs.

2. Determine (e.g. by a Pareto analysis or multicriteria methods) which products worth measuring risk (e.g. those more costly, those which generate more profitability).

3. Identify sources of uncertainty and characterize them.

4. Calculate the CaR.

As it has been mentioned before, Historical Simulations, Delta-Normal Approach and the Monte Carlo Simulation can be used for the computation of CaR. The calculation of CaR with historical simulation involves using different cost records for a product or service and then, construct a distribution of potential future cost of such products. The CaR will be calculated considering which values are greater or equal to a certain percentile, for example the 95th percentile. On the other hand, the normal delta approach is based on the assumption that the cost behavior fits a normal distribution. Based on this assumption it can be determined the probability distribution of the cost and then using standard mathematical properties of the normal distribution it is determined the cost that will equal or exceed a defined percentage, for example the percentage associated with CaR. Finally, the Monte Carlo Simulation method can be included in the experimental branch of mathematics that deals with experiments on random numbers based on statistical probabilities (Hammersley and Handscomb, 1964). The simulation can be

used to value certain instruments and to derive risk measures, when they are not available or when an analytical model is inapplicable. Essentially, the analyst describes, in a probabilistic sense, the variables that determine the value by their respective probability distributions. Then, using a random number generator, possible values for the variable under analysis are obtained (Nachtmann and Needy, 2003).

Thus, CaR, considering a pre-defined confidence level, takes into account the worst expected outcome in terms of cost in a certain period. Unlike VaR which considers the left tail of the distribution of output (i.e. lower income) CaR highlights the right part of the distribution (i.e. higher costs), as shown in Figure 1.

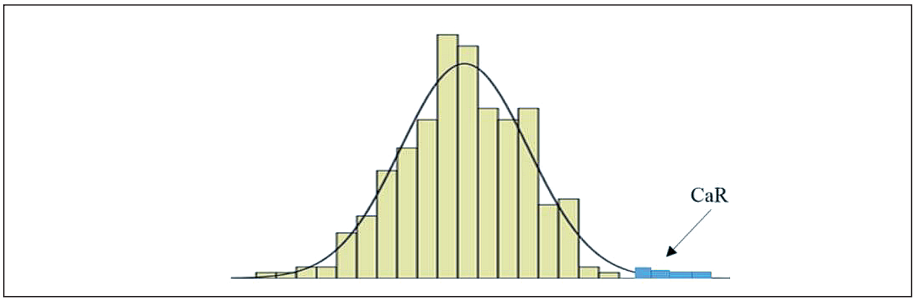


Figure 1. Costing at Risk.

3. PROBLEM AND COMPUTATIONAL RESULTS

A typical problem of manufacturing costs is used here to explain CaR. The case of Colombian coffee producers is particularly interesting in this context.

The manufacturing cost of goods and services refers to the sum of the value of resources used in the production or manufacture of a given quantity of a product or a service. It is presented in terms of raw materials, labor and general expenses or overhead incurred. Thus, manufacturing costs is composed by three main components as it is presented below.

$$\boldsymbol{C(x) = F1(x) + F2(x) + F3(x)} \tag{1}$$

The manufacturing cost can be represented also in two components instead of three: Prime Costs (i.e. raw materials and labor costs) plus Overhead or Raw Materials plus Conversion Costs (labor costs and overhead).

There are two important factors influencing the variability of the production cost of Colombian coffee. Firstly Colombian coffee is produced across the country, from north to south, creating a considerable cultural variability with impact not only in terms of labor aspects but also in terms of production methods and therefore cost. A second factor has to do with the existence of more than half a million of Colombian coffee growing families, 96% of them with less than 5 hectares of coffee plantations representing a huge variability in terms of production practices (Federations, 2015b). Coffee is a product that requires specific climatic conditions for growth. It is a unique product

of the tropics, its cultivation requires special conditions of terroir, i.e. soil, temperature, atmospheric precipitation and altitude above sea level. Typically, these features are not constant for all crops generating variability in the components of the manufacturing cost. Furthermore, the income of the coffee farmer is calculated using a price which depends from the stock price. In this context, it is necessary to have a good knowledge of the values of the different production costs and how the profitability of the producer is affected.

For the computational application data, they were used more than 200 simulated values for each of the three components of manufacturing cost – i.e. Materials, Direct Labor, and Overhead per kilogram produced. Different probability distributions have been constructed generating different associated risks to the computed production cost. To obtain these series it was took into account public information about coffee production costs available by the Ministry of Colombian Agriculture and existing information on the National Federation of Colombians Coffee Growers (AGRONET, 2015; Federations, 2015a).

For each of the four elements, material, labor, overhead costs and total cost per kilogram produced, an analysis in two steps of the data has been performed. Firstly, in an exploratory phase frequency histograms and skewness have been analyzed. The graphs of the obtained histograms are presented in Figure 2.

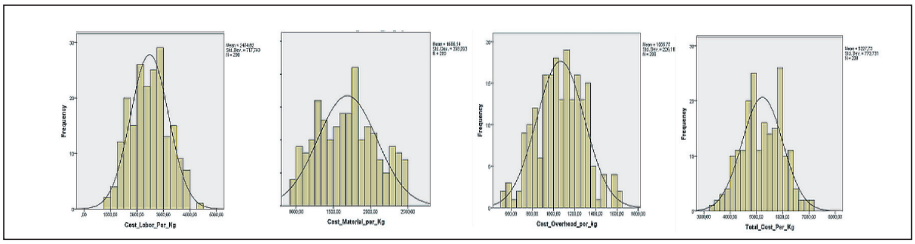


Figure 2. Histograms.

The second phase involves adjusting the series to a probability distribution, in this case qq-plot analysis and K-S test with the Lilliefors correction have been used. For data analysis was used SPSS 22. Table 1 and Figure 3 show the summary of the results which have been obtained.

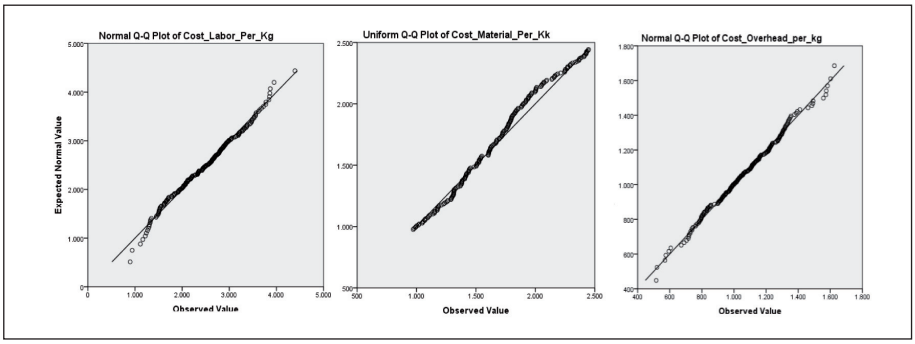


Figure 3. Q-Q plots for the Input Variables.

Kolmogorov-Smirnov Test						Kolmogorov-Smirnov Test For Uniform Distributions					
		Labor	Material	Overhead	Total Cost			Labor	Material	Overhead	Total Cost
N		200	200	200	200	N		200,00	200,00	200,00	200,00
Normal	Mean	2474,82	1686,14	1066,77	5227,73	Uniform	Minimum	896,72	971,94	515,80	3250,57
Parameters[a, b]	Std. Deviation	717,74	398,09	226,12	773,73	Parameters[a, b]	Maximum	4394,35	2447,23	1625,16	7046,71
Most Extreme Differences	Absolute	0,04	0,07	0,04	0,06	Most Extreme Differences	Absolute	0,18	0,09	0,17	0,17
	Positive	0,04	0,07	0,04	0,06		Positive	0,18	0,09	0,17	0,13
	Negative	-0,03	-0,05	-0,03	-0,06		Negative	-0,09	-0,05	-0,14	-0,17
TestStatistic		0,04	0,07	0,04	0,06	Kolmogorov-SmirnovZ		2,51	1,22	2,46	2,37
Asymp.Sig. (2-tailed)		,200[c,d]	,041[c]	,200[c,d]	,051[c]	Asymp.Sig. (2-tailed)		0,00	0,10	0,00	0,00
a. Test distribution is Normal. b. Calculated from data. c. Lilliefors Significance Correction. d. This is a lower bound of the true significance.						a. Test distribution is Uniform. b. Calculated from data.					

Table 1. K-S test for labor cost, material cost, overhead cost and total cost.

As shown in Table 1, the total cost per kilogram range between $3,250.57 and $7,046.71 Colombian pesos with an average of $5,227.73, and it is noteworthy that the highest average value of the parameters is given by the cost of labor per kg which is the cost element with higher variability for what it can be termed the most critical variable in terms of cost risk.

After defining the probability distributions, CaR has been calculated using historical simulation, the delta-normal, and Monte Carlo simulation. Table 2 shows the results obtained for CaR with 95% confidence using the three aforementioned methods. For

the Monte Carlo simulation it was used @risk. It is to highlight that Monte Carlo simulation allows to include correlations between the input variables and the variables that influence the variability of the total cost. Furthermore, via a sensitivity analysis it is possible to identify the critical variables or variables that should be under observation by managers and decision makers.

Methodology	Estimated CaR Historical
Simulations	$6.428,88
Delta-Normal Approach	$6.500,39
Monte Carlo Simulation	$6.530,61

Table 2. CaR Results.

In this case, the closeness of values may be due because the input variables are relatively reduced, and also because two of them could be adjusted to a normal distribution. Thus, with different probability distributions, more input variables and a different structure of the output equation it would be possible to obtain significant differences.

The three methods which have been presented are easy to understand and to systematize. The historical simulation has the disadvantage that does not permit incorporating elements associated with future knowledge and assumes that the past is replicable which is questionable in an economically changing environment. With respect to the normal approximation, it is possible that the distribution of the output variable may follow other distribution than normal. The Monte-Carlo method is a flexible methodology that allows the inclusion of correlations between variables and accepts parameters of the distributions which are not exactly known.

4. CONCLUSIONS

The results of CaR allow the identification of the most critical component, i.e. the component that most influence the variability of cost. From the point of view of managers, it could be a component on which more control should be exercised. Furthermore, the identification of potential losses related to excessive costs may permit to propose intervention measures or strategies of risk mitigation. Thus, measuring the uncertainty associated with costing systems will turn cost management

more effective through for example a proper budgeting and a better financial risks management.

CaR takes into consideration the worst expected outcome in terms of cost in a certain period and can be used to measure risk in costing systems. Particularly production costs may vary considerably and the reasons behind such variability may be not well understood. This paper proposed a methodology to measure risk in this context. Further work can be made to estimate CaR with other parametric, non-parametric and semi-parametric methods, in addition to applying this concept in other crops, such as growing grapes for winemaking and in the health care sector.

REFERENCES

AGRONET. (2015). *Sistema de Estadísticas Agropecuarias, SEA*. Retrieved from http://www.agronet.gov.co/agronetweb1/Estad%C3%ADsticas.aspx

Donaldson, L. (2001). *The Contingency Theory of Organizations*. Sage.

Esmalifalak, H., Albin, M. S., and Behzadpoor, M. (2014). "A Comparative Study on the Activity Based Costing Systems: Traditional, Fuzzy and Monte Carlo Approaches". *Health Policy and Technology*, 1-10. doi:10.1016/j.hlpt.2014.10.010

Federations, C.C.G. (2015a). *Historical Statistics*. Retrieved from http://www.federaciondecafeteros.org/particulares/en/quienes_somos/ 119_estadisticas_historicas/

Federations, C.C.G. (2015b). *Our Coffee Growers*. Retrieved from http://www.federaciondecafeteros.org/particulares/es/nuestros_caficultores

Hammersley, J.M., and Handscomb, D.C. (1964). *Monte Carlo Methods*. Methuen & CO LTD, Ed. (vol. 1). London: Springer.

Nachtmann, H., and Needy, K.L. (2003). "Methods for Handling Uncertainty in Activity Based Costing Systems". *The Engineering Economist*, *48*(3), 259-282. doi: 10.1080/00137910308965065

52. ON RISK PROCESS WITH PARETO MIXING

P. Jordanova[1]
Shumen University, Faculty of Mathematics and Informatics, Bulgaria,
pavlina_kj@abv.bg

Y. Nefedova
Lomonosov Moscow State University, FCMC, Russia,
y.nefedova@gmail.com

M. Stehlík
Johannes Kepler University in Linz, Dep. of Appl. Stat., Austria, Milan.
Stehlik@jku.at.
Universidad Técnica Federico Santa María, Valparaíso,
Department of Statistics, Chile.

ABSTRACT

Risk processes with mixed exponentially-Pareto distributed inter-arrival times are considered. The number of claims is mixed Poisson-Pareto distributed. Several important approximations of these risk processes for different types of claim sizes are obtained and the convergence rate in the case of mixed diffusion approximation is investigated.

1. INTRODUCTION

Minimization of probability of ruin has become a classical criterion in risk optimization and portfolio management. In recent years it has been extensively studied being

[1] The first author is supported by Grant No. RD-08-265/10.03.2015 of Shumen University, Bulgaria.

of a natural interest to the policy-makers and supervision of the insurance companies. In a long horizon approximations of the risk process (prs) could be very useful. Such setup can be interesting e.g. for Basel II framework. Iglehart (1969) obtained diffusion approximation in the Cramer-Lundberg model for claims with finite variance. Furrer *et al.* (1997) derived α-stable approximation for the renewal risk model with existing expectation of inter arrival times, and non-existing variance, $\alpha \in (1, 2)$. Jordanova (2005b) obtained α-stable approximation in analogous model for $\alpha \in (0, 1)$. Grandel (1991) generalizes the idea about diffusion approximation, in the case when the counting prs is a Cox-prs. Approximations of mixed risk processes are considered also in Jordanova (2005) and Pancheva *et al.* (2007). The last derived upper and lower bounds for ruin probability under assumption that sum of claims weakly converges to some α-stable Lévy motion. Pancheva *et al.* (2011) made survey of all mentioned results and generalized them for a particular case of non-identically distributed claims. It is worth noting that the dependence between the interarrival times is related to the Archimedean survival copulas as discussed in Albrecher *et al.* (2011). In this sense, it is possible to obtain some explicit formula for the ruin probability by conditioning arguments.

Stochastic interest rates play an important role considering long maturity contracts. This is well-known in the literature, see e.g. Bacinello *et al.* (2000), Suárez-Taboada and Vázquez (2012) or Pavlenko (2000) and references therein. For a specific classification according to monotonicity see Stehlík *et al.* (2015).

In this paper we consider random time transformed Cramer-Lundberg model with Pareto mixing variable(v.) and obtain corresponding approximations of the risk process. Both random and non-random interest rates are discussed.

2. DESCRIPTION OF THE MODEL AND PRELIMINARY RESULTS

Denote by $\Gamma(x, t) = \int_t^\infty y^{x-1}e^{-y}\,dy$ the upper incomplete gamma function and by $E_p(z) = z^{p-1}\Gamma(1-p, z), z \in R, p \in R$, the generalized exponential integral. Some of their properties could be found in Milgam (1985) or Olver *et al.* (2010). Let $\alpha > 0$ and $\delta > 0$. Consider the following insurance model:

(A1) For all $n \in N$ the inter-arrival times $(Y_1, Y_2, \ldots Y_n)$ constitute an Exponentially-Pareto distributed random (r.) vector with parameters α and δ, i.e.

$$P(Y_1 > x_1, ..., Y_n > x_n) = \alpha E_{\alpha+1}((x_1 + ... + x_n)\delta), x_1 > 0, \ ..., \ x_n > 0. \quad (1)$$

This distribution is introduced and partially investigated in Jordanova *et al.* (2013). There is shown that Y_i, $i = 1, 2, \ldots$ are Exp.-Pareto distributed r. variables with parameters α and δ. The last implies that $P(Y_1 < \infty) = 1$,

$$EY_i = \frac{\alpha}{\delta(1+\alpha)}, \quad Var\ Y_i = \frac{\alpha(2+2\alpha+\alpha^2)}{\delta^2(\alpha+1)^2(\alpha+2)}, \quad i = 1, 2, \ldots \quad (2)$$

(A2) $T_0 = 0$ and denote by $T_n = Y_1 + \ldots + Y_n$, $n = 0, 1, \ldots$ the claim arrival times in the insurance company. Then T_n is Erlang-Pareto distributed with parameters α, δ and $n \in N$.

(A3) Consider a time-space point process,

$$N = \{(T_k, X_k) : k = 0, 1, ...\} = \{(Y_0 + ... + Y_k, X_k) : k = 0, 1, ...\}, \quad X_0 = 0$$

with independent time and space components. Here X1, X2, ... are i.i.d. r.vs with d.f. F, mean $EX_1 = a \leq \infty$ and variance $Var\ X_1 = \sigma^2 \leq \infty$. They represent the corresponding clime sizes.

(A4) Define the counting prs $\{N^*(t) = max\{k \geq 0 : T_k \leq t\} : t \geq 0\}$. For all fixed $t > 0$ it counts the number of claims in $(0, t]$.

Jordanova *et al.* (2013) show that it is a homogeneous in time Mixed Poisson prs with Pareto mixing v. It is not a renewal prs. For all fixed $t > 0$ $N^*(t)$ is a Mixed Poisson-Pareto distributed r.v. with parameters α and δt. This distribution is proper, therefore for all $t < \infty$, $N^*(t)$ is a.s. finite. $\{N^*(t): t \geq 0\}$ has the following representation $N^*(t) = N(\Lambda_{\alpha,\delta}\, t)$, $t \geq 0$, where $\Lambda_{\alpha,\delta}$ is Pareto distributed r.v. with distribution function (d.f.) $P(\Lambda_{\alpha,\delta} < x) = 1 - \frac{\delta^\alpha}{x^\alpha}, x > \delta$. and N is a homogeneous Poisson prs with intensity 1, independent on $\Lambda_{\alpha,\delta}$. Therefore

$$P(Y_1 > x_1, ..., Y_n > x_n) = \alpha E_{\alpha+1}((x_1 + ... + x_n)\delta), x_1 > 0, \ ..., \ x_n > 0. \quad (3)$$

It is easy to obtain that

$$\frac{EN^*(t)}{t} = \frac{\alpha\delta}{(\alpha-1)} = E\Lambda_{\alpha,\delta}, \quad \alpha > 1, \tag{4}$$

$$Var\frac{N^*(t)}{t} = \frac{\alpha\delta}{t(\alpha-1)} + \frac{\alpha\delta^2}{(\alpha-2)} - \frac{\alpha^2\delta^2}{(\alpha-1)^2} = \frac{E\Lambda_{\alpha,\delta}}{t} + Var\Lambda_{\alpha,\delta}, \quad \alpha > 2. \tag{5}$$

$$\frac{N^*(t)}{t} \to \Lambda_{\alpha,\delta},\ a.s.\ t \to \infty,$$

$$\frac{N^*(t) - EN^*(t)}{\sqrt{Var\, N^*(t)}} \to \frac{\Lambda_{\alpha,\delta} - E\Lambda_{\alpha,\delta}}{\sqrt{Var\ \Lambda_{\alpha,\delta}}},\ a.s.,\ t \to \infty, \quad \alpha > 2. \tag{6}$$

By T 6.1.2 of Rolski (2009) and continuity of the composition

$$\frac{N^*(nt) - \Lambda_{\alpha,\delta} nt}{\sqrt{n}} \Longrightarrow W(\Lambda_{\alpha,\delta} t),\ n \to \infty, \tag{7}$$

Here and further on $\{W(t), t \geq 0\}$ stands for a Standard Brownian motion and $\to$ denotes convergence in J_1 topology (Bingham (1971)). We can apply Donsker Invariance Principle (Billingsley, 1977) for $T_i - ET_i$, $i = 1, 2, \ldots$ and continuity of the composition, Th. 3.1. W.Whitt (1980) and obtain that

$$\frac{T_{N^*(nt)} - ET_{N^*(nt)}}{\sqrt{n}} \Longrightarrow W(\Lambda_{\alpha,\delta} t) = \Lambda_{\alpha,\delta}^{1/2} W(t), \quad n \to \infty.$$

(A5) Denote by S(t), $t \geq 0$ the total claim amount process up to time t, i.e.

$$S(t) = \sum_{k=1}^{N^*(t)} X_k,\ \sum_{k=1}^{0} = 0.$$

When $EX_1 < \infty$ by Wald equality $ES(t) = EN^*(t)EX_1$.

- If $E|X_1|p < \infty$ for some $p \in (0, 2]$ by Th 2.5.5 Embrechts *et al.* (1997) and (6) we obtain that:

  – for $p \in (0,1)$ $\frac{S(nt)}{\sqrt[p]{N^*(nt)}} \sim \frac{S(nt)}{\sqrt[p]{nt\Lambda_{\alpha,\delta}}} \to 0,\ n \to \infty$, a.s.

  Moreover if $F \in DA(S_\beta)$ for some $\beta \in (0, 1]$ and $DX_1 = \infty$, by Meerschaert *et al.* (2004) and Th. 3.1. Whitt (1980)

$$\frac{S(nt)}{F^{\leftarrow}(1-\frac{1}{n})} \Longrightarrow S_{\beta}(\Lambda_{\alpha,\delta}t),\ n \to \infty, \tag{8}$$

where S_β is an one sided β-stable Lévy motion with $Ee^{-yS_\beta(t)} = e^{-\frac{ty^\beta}{\Gamma(1-\beta)}}$ and $F \leftarrow$ denotes the inverse function to F. Here and further on ~ means asymptotic equality in distribution.

– for p ∈ [1, 2) a.s.

$$\frac{S(nt)-N^*(nt)EX_1}{\sqrt[p]{N^*(nt)}} \sim \frac{S(nt)-nt\Lambda_{\alpha,\delta}EX_1}{\sqrt[p]{N^*(nt)}} \sim \frac{S(nt)-nt\Lambda_{\alpha,\delta}EX_1}{\sqrt[p]{nt\Lambda_{\alpha,\delta}}} \to 0,\ n \to \infty.$$

In particular

$$\frac{S(t)}{N^*(t)} \to EX_1,\ a.s.\ t \to \infty, \tag{9}$$

$$\frac{S(t)}{t} \to \Lambda_{\alpha,\delta}EX_1,\ a.s.\ t \to \infty. \tag{10}$$

If $F \in \mathrm{DA}(S_\beta)$ for some β ∈ (1, 2] and $DX_1 = \infty$, by Furer *et al.* (1997) and (6) we obtain that:

$$\frac{S(nt) - nt\Lambda_{\alpha\delta}EX_1}{F^{\leftarrow}(1-\frac{1}{n})} \sim \frac{S(nt) - N^*(nt)EX_1}{F^{\leftarrow}(1-\frac{1}{n})} \Longrightarrow S_\beta(\Lambda_{\alpha,\delta}t), \tag{11}$$

where S_β is an one sided β-stable Lévy motion with

$$S_\beta(1) \sim S_\beta(-\beta\Gamma(-\beta)cos\frac{\pi\beta}{2}, 1, 0) = S_\beta(\Gamma(1-\beta)cos\frac{\pi\beta}{2}, 1, 0). \tag{12}$$

– If $VarX_1 < \infty$ by Functional Central Limit Th. (Billingsley (1977)), continuity of the composition and (6), we obtain that for $n \to \infty$,

$$\frac{S(nt) - N^*(nt)EX_1}{\sqrt{nVar\,X_1}} \sim \frac{S(nt) - nt\Lambda_{\alpha,\delta}EX_1}{\sqrt{nVar\,X_1}} \Longrightarrow W(\Lambda_{\alpha,\delta}t). \tag{13}$$

It is well known that the random processes $W(\Lambda_{\alpha,\delta}\cdot)$ and $\sqrt{\Lambda_{\alpha,\delta}}W(\cdot$ are stochastically equivalent.

3. THE RISK PROCESS AND ITS APPROXIMATION

In this section we will consider the risk reserve process $R(t)$, $t \geq 0$, related to the model described in (A1) - (A5).

(A6) Let $R(t) = c(t) - S(t)$, $t \geq 0$, where $c(t)$ is the income process. The initial capital $c(0) = u$ is non-random and $c(t) - c(0)$ is the sum of premiums up to time t.

Then $ER(t) = Ec(t) - EN^*(t)EX_1 = Ec(t) - \frac{\alpha\delta t EX_1}{\alpha-1}$, $t \geq 0$, $\alpha > 1$. Consider the net profit condition $\lim_{t\to\infty} \frac{Ec(t)}{t} > \frac{\alpha\delta EX_1}{\alpha-1}$. In particular, for $c(t) = u + ct$ and $EX_1 = \frac{1}{\mu}$, it becomes $c > \frac{\alpha\delta}{\mu(\alpha-1)}$ and the safety loading is $\rho = c - \frac{\alpha\delta}{\mu(\alpha-1)}$. If $X_1, X_2, \ldots$ are independent exponentially distributed with $EX_1 = \frac{1}{\mu}$, following Albrecher *et al.* (2011) we could obtain the explicit form for the probability of ruin in infinite time. In many cases these explicit formulae is impossible to find, e.g. for Pareto claim sizes. In such cases the following approximations of the risk processes could be useful.

- Case 1. Suppose $VarX_1 = \sigma^2 < \infty$. We can transform the initial point process in such a way that the corresponding sequence of risk processes to converge. Consider the point process

$$N_n^* = \{(\frac{T_k}{n}, \frac{X_k - EX_1}{\sqrt{n\sigma^2}}) : \ k = 1, 2, \ldots\}.$$

  Then under condition of existence of deterministic, non-negative, strictly in creasing function (d.n.s.i.f.) $c_0(t)$ such that $\frac{c(nt)-\Lambda_{\alpha,\delta} ntEX_1}{\sqrt{n\sigma^2}} \Longrightarrow c_0(t)$, by (13) the sequence of the risk processes

$$R_n(t) = \frac{R(nt)}{\sqrt{n\sigma^2}} \Longrightarrow c_0(t) - W(\Lambda_{\alpha,\delta} t) = c_0(t) - \Lambda_{\alpha,\delta}^{1/2} W(t).$$

- Case 2. Suppose $VarX_1 = \infty$, $EX_1 < \infty$, $\overline{F}(x) := P(X_1 \geq x) \in RV_{-\beta}$ $\beta \in (1, 2)$ and $\frac{c(nt)-\Lambda_{\alpha,\delta} n.t.EX_1}{F^{\leftarrow}(1-\frac{1}{n})} \Longrightarrow c_0(t)$, where $c_0(t)$ is a d.n.s.i.f. In view of (11) we obtain that the sequence of risk processes

$$R_n(t) = \frac{R(nt)}{F^{\leftarrow}(1-\frac{1}{n})} \Longrightarrow c_0(t) - S_\beta(\Lambda_{\alpha,\delta} t) = c_0(t) - \Lambda_{\alpha,\delta}^{1/\beta} S_\beta(t),$$

where S_β is the one-sided alpha stable Lévy motion, defined in (12). In this case one possible choice of $c_0(t)$ is $c_0(t) = u_0 + t^{1/\beta}$. It corresponds to $u + F^{\leftarrow}(1 - \frac{1}{t}) + \Lambda_{\alpha,\delta} t E X_1$.

- Case 3. Let $F(x) := P(X_1 < x)$ and $\overline{F} \in RV_{-\beta}$, $\beta \in (0,1)$. Then $EX_1 = \infty$. Then we do not need centering in order to obtain non-degenerate limit process.

  The net profit condition is replaced by the condition that there exists a d.n.s.i.f. $c_0(t)$ such that $\lim_{n\to\infty} \frac{c(nt)}{F^{\leftarrow}(1-\frac{1}{n})} = c_0(t)$, $t \geq 0$. Under this condition

$$R_n(t) = \frac{R(nt)}{F^{\leftarrow}(1 - \frac{1}{n})} \Longrightarrow c_0(t) - \Lambda_{\alpha,\delta}^{1/\beta} S_\beta(t),$$

  where $S\beta(t)$ is one-sided β-stable Levy motion, described in (8).

REFERENCES

Albrecher, H., Constantinescu, C., and Loisel, S. (2011). "Explicit Ruin For Mulas for Models with Dependence among Risks". *Insurance: Mathematics and Economics*, 48, 2, 265-270.

Bacinello, A.R., and Persson, S.A. (2000). "Design and Pricing of Equity-Linked Life Insurance Under Stochastic Interest Rates". *EFA 2001 Barcelona Meetings*. http://ssrn.com/abstract=276701 or http://dx.doi.org/10.2139/ssrn.276701.

Billingsley, P. (1977). *Convergence of Probability Measures*, John Wiley and Sons.

Embrechts, P., Klueppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events: For Insurance and Finance.*

Furrer, H., Michna, Z., and Weron, A. (1997). "Stable Lévy Motion Approximation in Collective Risk Theory". *Insurance: Mathematics and Economics*, 20, 97-114.

Iglehart, D.L. (1969). "Diffusion Approximations in Collective Risk Theory, *J. Appl. Probab.*, 6, 295-292.

Jordanova, P., (2005). "Compound α-stable Approximation for a Total Claim Ammount Process under Condition of Mixed Poisson Claim Arrival Process". *Proc. Russe University*, Bulgaria, 44, 6.1, 32-37 (in Bulgarian).

Jordanova, P. (2005b). "Compound α-stable Approximation for a Risk Process". *Proceedings of International Scientific Conference UNITECH'05, Bulgaria*, Technical University of Gabrovo, III, 403-407 (in Bulgarian).

Jordanova, P., Dušek, J., and Stehlík, M. (2013). "Microergodicity Effects on Ebullition of Methane Modelled by Mixed Poisson Process with Pareto Mixing Variable". *Chemometrics and Intelligent Laboratory Systems*, 128, 124-134. To Cox processes. *Journal of Mathematical Sciences*, 81, 5, 2951-2956.

Meerschaert, M.M., and Scheffler, H.P. (2004). "Limit Theorems for Continuous-Time Random Walks with Infinite Mean Waiting Times". *J. Appl. Probab.*, 41 (3), 623-638.

Milgam, M.S. (1985). "The Generalized Integro-Exponential Function". *Math. Comput.*, 44, 443-458.

Olver, F.W.J., Lozier, D.W., Boisvert, R.F., and Clark, C.W. (2010). *NIST Handbook on Mathematical Functions*, Cambrige University Press.

Pancheva, E., Kolkovska, E., and Jordanova, P. (2007). "Random Time-Changed Extremal Processes". *Theory Prob. Appl.*, 51, 4, 645-662.

Pancheva, E., and Mitov, I. (2011). "Approximation of the Risk Process. A Survey". *Mathematica Balkanica*, 25, 3, 307-316.

Pavlenko, O.I. (2000). "Risk Processes with Random Interest Rate". *Cybernetics and Systems Analysis*, 36, 5, 743-748.

Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (2009). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons.

Stehlík, M., Potocký, R., Kiselák, J., and Jordanova, P. (2015). "On Generalized Interest Rate Dynamics", accepted it *Appl. Math. Inf. Sci*.

Suárez-Taboada, M., and Vázquez, C., (2012). "Numerical Solution of a PDE Model for a Ratchet-Cap Pricing with BGM Interest Rate Dynamics". *Applied Mathematics and Computation*. 218, 9, 5217-5230.

Whitt, W. (1980). "Some Useful Functions for Functional Limit Theorems". *Mathematics of Operations Research*. 5, 1, 67-85.

53. ON THE MICHAELIS-MENTEN MODEL

C.P. Kitsos and T.L. Toulias
Technological Educational Institute of Athens, Informatics Department, Greece,
{xkitsos, t.toulias}@teiath.gr

ABSTRACT

The target of this paper is to discuss the well-known Michaelis-Menten model, which plays an important role in pharmacokinetics, from a fitting as well as from an optimal design point of view. We prove that the M-M model has not unique parameter estimation, through the nonlinear least squares method, and hence there is not a unique optimal experimental design.

1. THE MICHAELIS–MENTEN MODEL

A general theory for enzyme kinetics was firstly developed by Michaelis and Menten (1913) in their pioneering work, where the metabolism of an agent is described by a reaction rate. For a compact review of Bioassays see Kitsos (2012). The basic toxic kinetic model of metabolism is a Michaelis-Menten (M-M) model. This is discussed very briefly: When an enzyme, say E, is combined reversibly with a substrate, say S, to form an enzyme-substrate complex, say ES, which can be dissociate or proceed to the product, say P, the enzyme-substrate reaction scheme $E + S \underset{k_1}{\overset{k_2}{\leftrightarrows}} ES \xrightarrow{k_3} E + P$ is assumed, with *k1, k2* and *k3* being the associated rate constants. We let $K_{\mathrm{M}} := (K_2 + k_3)/k_1$, known as the M-M constant, and $\mathrm{V}_{\max} := k_3 C_{\mathrm{T}}$, where CT is the total enzyme concentration. Then, a plot of the initial velocity of reaction V against the concentration of substrate C_{S}, will provide the M–M rectangular hyperbola of the form

$$V = f(C_{\mathrm{S}}) = f(C_{\mathrm{S}};\, \boldsymbol{\theta}) := \frac{V_{\max} C_{\mathrm{S}}}{K_{\mathrm{M}} + C_{\mathrm{S}}}, \tag{1}$$

where the parameter vector $\theta = (\theta_1, \theta_2) := (\mathrm{V}_{\max}, K_{\mathrm{M}})$. In practice, we only have readings of the form (1) with $f(x; \theta) = u$ with $x := C_{\mathrm{S}}$. That is, only the stochastic model of the form $y_i = u_i + e_i$, $i = 1, 2, \ldots,$ n is obtained, as the readings y_i are associated with noise. In the sequence we shall refer to u as the expected value of y, i.e. $\mathrm{E}[y] = u$.

Different linear transformations have been suggested, Currie (1982), Dowd and Riggs (1965), to estimate (with a linear regression fit) the involved parameters, as the input variable C_{S} and the response V are curvilinearly related. Namely, the Eadie-Hofstee, the Hanes-Wolf and the Lineweaver-Burk (or "double reciprocal") linearizations. These linear transformations are given, respectively, by

$$V = V_{\max} - K_{\mathrm{M}}\frac{V}{C_{\mathrm{S}}}, \quad \frac{C_{\mathrm{S}}}{V} = \frac{K_{\mathrm{M}}}{V_{\max}} + \frac{1}{V_{\max}}C_{\mathrm{S}} \text{ and } \frac{1}{V} = \frac{1}{V_{\max}} + \frac{K_{\mathrm{M}}}{V_{\max}}\frac{1}{C_{\mathrm{S}}}, \tag{2}$$

with the last one (double reciprocal) being the most popular. The Hanes-Wolf linearization has been shown in a very early work by Dowd and Riggs (1965), as the most efficient. Endvenyi and Chan (1981), Currie (1982) discussed the heteroscedasticity in the M-M model, while Gilberg *et al.* (1999) provided an extensive discussion. Endvenyi and Chan (1981) assumed that the error variance is proportional to the mean. To avoid heteroscedasticity problems in the "linearized" models (2) we should suggest to solve the model's Normal Equations and get the least square estimators when n readings of V and C_{S} are given, i.e. to solve

$$\sum_{i=1}^{n}\left(V_i - \frac{V_{\max}C_{\mathrm{S},i}}{K_{\mathrm{M}} + C_{\mathrm{S},i}}\right)\nabla V(C_{\mathrm{S},i}) = 0, \tag{3}$$

with V as in (1) and

$$\nabla V := \left(\frac{\partial V}{\partial V_{\max}}, \frac{\partial V}{\partial K_{\mathrm{M}}}\right)^{\mathrm{T}} = \left(\frac{C_{\mathrm{S}}}{K_{\mathrm{M}} + C_{\mathrm{S}}}, -\frac{V_{\max}K_{\mathrm{M}}C_{\mathrm{S}}}{(K_{\mathrm{M}} + C_{\mathrm{S}})^2}\right)^{\mathrm{T}}. \tag{4}$$

Then, the two Normal Equations can be easily obtained and solved numerically so that the estimates $\hat{\boldsymbol{\theta}} = (\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$ are obtained. For a single observation the Fisher's information matrix is $(\nabla \mathrm{V})^{\mathrm{T}}(\nabla \mathrm{V})$. Therefore, the average–per–information matrix $\mathrm{M}(\theta, \xi)$ is evaluated as

$$\sigma^{-2}n\mathbf{M}(\theta, \xi) = \sum_{i=1}^{n}\begin{pmatrix} C_{\mathrm{S},i}^2\tau_i^2 & -V_{\max}C_{\mathrm{S},i}^2\tau_i^3 \\ -V_{\max}C_{\mathrm{S},i}^2\tau_i^3 & V_{\max}C_{\mathrm{S},i}^2\tau_i^4 \end{pmatrix}, \tag{5}$$

with $\tau_i = 1/(K_M + C_S, i)$, $i = 1, 2, \ldots, n$. Thus, the 2 × 2 variance–covariance matrix is approximately equal to

$$\mathbf{C} = \mathbf{C}(\hat{\theta}, \xi) = \left(n\mathbf{M}(\hat{\theta}, \xi)\right)^{-1}. \tag{6}$$

Hence, we can derive asymptotic approximate confidence intervals for the involved parameters.

2. OPTIMAL DESIGN THEORY FOR THE M-M MODEL

We are forming the D–optimal design problem for the M-M model, Kitsos (2001), in the following steps:

Step 1. The D-optimal design for the M-M model, as in (1), does not depend on V_{max}. Indeed, it is

$$\nabla V = \begin{pmatrix} 1 & 0 \\ 0 & V_{\max} \end{pmatrix} \begin{pmatrix} C_S(K_M + C_S)^{-1} \\ -C_S(K_M + C_S)^{-2} \end{pmatrix}, \tag{7}$$

with ∇V as in (4). That is, the gradient vector ∇V can be written as a product of a matrix with elements only 0, 1, V_{max} and a vector which depends only on K_M. Then, following Hill (1980), the model is partially nonlinear and the parameter V_{max} does not influence the D-optimal design for estimating θ.

Step 2. The D-optimal design for the extended M-M model of the form

$$u := \frac{V_{\max} C_S}{K_M + C_S} + \theta_0 C_S, \tag{8}$$

for estimating $\theta = (V_{max}, K_M\theta_0)$ does not depend on V_{max} and θ_0. Moreover, the D_S-optimal design for estimating θ_0 depends on K_M. Indeed, consider

$$\sigma^{-2} n\mathbf{M}(\theta, \xi) = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \tag{9}$$

Let $\theta_1 = (V_{max}, K_M)$, $\theta_2 = \theta_0$, i.e. $\theta = (\theta_1, \theta_2)$. Then $M(\theta, \xi) = M(\theta_1, \xi)$ does not depend on $\theta_2 = \theta_0$. Hence, the D-optimal design depends only on θ_1. However, it eventually depends only on K_M, as

$$\nabla V = \mathrm{diag}(1, V_{\max}, 1) C_{\mathrm{S}} \left((K_{\mathrm{M}} + C_{\mathrm{S}})^{-1}, (K_{\mathrm{M}} + C_{\mathrm{S}})^{-2}, 1 \right)^{\mathrm{T}}, \tag{10}$$

with diag(C_{11}, C_{22}, C_{33}) being a 3 × 3 matrix with C_{ii}, $i = 1, 2, 3$ given and C_{ij}, $i = j = 1, 2, 3$ are equal to zero.

Step 3. If $C_{\mathrm{S}} \in (0, C_{\mathrm{U}}]$, the locally D-optimal design at $K_{\mathrm{M}} = K_0$ allocates half observations at points V and optimal concentration $C_{\mathrm{S}}^{\mathrm{opt}}$, where

$$C_{\mathrm{S}}^{\mathrm{opt}} = \frac{K_0 C_{\mathrm{U}}}{2K_0 + U}, \tag{11}$$

with C_{U} being the maximum allowable substrate concentration. If $C_{\mathrm{U}} >> K_0$ the locally D-optimal design ξ is

$$\xi^* = \begin{pmatrix} C_{\mathrm{U}} & K_0 \\ 0.5 & 0.5 \end{pmatrix}. \tag{12}$$

Indeed, this is true from (11).

Step 4. The corresponding value of the determinant of the D-optimal design is

$$d = \frac{V_{\max}^2 C_{\mathrm{U}}^6}{16 K_0^2 (K_0 + C_{\mathrm{U}})^6}. \tag{13}$$

See also Endrenyi and Chan (1981). If $C_{\mathrm{S}} \in [C_{\mathrm{L}}, C_{\mathrm{SU}}]$ then the optimum C_{S} is, through (11), at

$$C_{\mathrm{S}}^{\mathrm{opt}} = \frac{K_0 C_{\mathrm{U}}}{2K_0 + (C_{\mathrm{U}} - C_{\mathrm{L}})}, \quad K_0 > 0, \quad 0 < C_{\mathrm{L}} < C_{\mathrm{U}}. \tag{14}$$

Moreover, when n is odd, i.e. $n = 2q + 1$, we allocate q and $q + 1$ observations at the optimum design points so that two solutions can be obtained. With n large the experimenter decides where he will allocate the $q + 1$ observations.

Step 5. The D-optimal designs for estimating the parameter vector ($V_{\max}$, K_{M}), as in Step 3, are the D-optimal designs for estimating the ratio $\varphi = V_{\max} / K_{\mathrm{M}}$. Indeed, it is

$$V = \frac{V_{\max} C_{\mathrm{S}}}{K_{\mathrm{M}} + C_{\mathrm{S}}} = \frac{\frac{V_{\max}}{K_{\mathrm{M}}} C_{\mathrm{S}}}{1 + \frac{C_{\mathrm{S}}}{K_{\mathrm{M}}}} = \frac{\varphi C_{\mathrm{S}}}{1 + \varphi_0 C_{\mathrm{S}}}. \tag{15}$$

Thus, from (15) and Step 1, only prior guess for φ_0, i.e. for K_M, is needed to construct a D-optimal design as in (12) or (11), replacing K_0 with φ_0. If we let $\beta = (\varphi, \varphi_0)$ as in (15) and $\theta = (V_{max}, K_M)$ as in (1), then the variance of $\hat{\beta}$ is associated to $\mathrm{Var}(\hat{\beta}) \cong (\nabla\beta)^{\mathrm{T}}\mathbf{M}^{-1}(\theta,\xi)(\nabla\beta)$. .

Step 6. Consider the D-optimal design (12) with $n = 2$, i.e. we allocate one observation at each optimal point. Then, the asymptotic covariance matrix is

$$\left(2\sigma^{-2}\mathbf{M}(\theta_0/2)\right)^{-1} = \begin{pmatrix} 1 & 2\omega \\ 2\omega & 20\omega^2 \end{pmatrix}, \tag{16}$$

with $\theta_0 = (V_{max}, K_0)$ and $\omega = K_0/V_{max}$. Indeed, (5) implies, due to (12), that

$C_{S,1} = C_U$, and therefore $\tau_1 = \frac{1}{K_0+C_U} = \frac{1}{U(K_0/C_U+1)} = \frac{1}{C_U}$, $\tau_1^2 = \frac{1}{C_U^2}$,

$C_{S,2} = K_0$, and therefore $\tau_2 = \frac{1}{K_0+K_0} = \frac{1}{2K_0}$.

With some matrix algebra we evaluate

$$2\sigma^{-2}\mathbf{M}(\theta_0/2) = \begin{pmatrix} 5/4 & -\varphi/8 \\ -\varphi/8 & \varphi^2/16 \end{pmatrix}, \tag{17}$$

with $\varphi = V_{max}/K_0 = 1/\omega$. From (17) it is easy to calculate that the determinant of the matrix M equals $\varphi^2/16$ and thus (16) holds. Notice that the value of the determinant $\varphi^2/16$ coincides with the value of (13) under the (12) design.

3. NONLINEAR LSE FOR THE M-M MODEL

An analytical method for the Nonlinear Least Square (NLLS) estimation of the M-M model is proposed and discussed here. In particular, the following Theorem provides a compact analytic methodology for the "actual" NLLS estimation (i.e. not iterative algorithms are involved) of the estimation of the M-M model's parameters.

Recall that the (residuals) sum of squares $S = S(\boldsymbol{\theta}) := \sum_{i=1}^{n}[V_i - f(x_i;\boldsymbol{\theta})]^2$, where $(x_i) := (C_{S,i}) \in \mathbb{R}^n$, $n > 2$, and $(V_i) \in \mathbb{R}^n$ are the readings' vectors for the concentration and the reaction's velocity. The estimated parameters $\hat{\theta}_1 = \hat{V}_{max}$ and $\hat{\theta}_2 = \hat{K}_M$, from

the estimated vector $\hat{\boldsymbol{\theta}}$, are obtained when $S(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}\in\Theta}\{S(\boldsymbol{\theta})\}$, i.e. when ∇S = 0, or equivalently, when the normal equations (3) are satisfied.

Theorem 3.1. The NLLS estimators $\hat{V}_{\max}$ and $\hat{K}_{\mathrm{M}}$ of the M–M model are the ones among the $K < 4n - 5$ possible estimators $\hat{\boldsymbol{\theta}}_k = (\hat{\theta}_1^k, \hat{\theta}_2^k)$, which provide 1 2 the minimum sum of squares, i.e. $S(\hat{\boldsymbol{\theta}}) = S(\hat{V}_{\max}, \hat{K}_{\mathrm{M}}) = \min\{S(\hat{\boldsymbol{\theta}}_k)\}_{k=1,2,\ldots,K}$, where

$$\hat{\theta}_1^k = \left(\sum_{i=1}^{n} \frac{u_i C_{\mathrm{S},i}}{\hat{\theta}_2^k + u_i}\right)\left[\sum_{i=1}^{n}\left(\frac{C_{\mathrm{S},i}}{\hat{\theta}_2^k + C_{\mathrm{S},i}}\right)^2\right]^{-1}, \quad k = 1, 2, \ldots, 4n-5, \tag{18}$$

provided that $\hat{\theta}_2^k \neq -C_{\mathrm{S},i}, -V_i,\ i = 1, 2, \ldots, n$, while $\hat{\theta}_2^k$ are $K < 4n - 5$ real roots of the following (4n – 5)–degree polynomial of ϑ:

$$P(\vartheta) := \sum_{\substack{i,j=1\\(i\neq j)}}^{n} u_i C_{\mathrm{S},i} C_{\mathrm{S},j} (C_{\mathrm{S},j} - C_{\mathrm{S},i})(\vartheta + C_{\mathrm{S},i})^2 (\theta_2 + C_{\mathrm{S},j}) \prod_{(i,j\neq)k=1}^{n} (\vartheta + C_{\mathrm{S},k})^4.$$

Proof. Solving the first normal equation (3) in terms of $\mathrm{V}_{\max}$ and substituting to the second we obtain

$$\left[\sum_{i=1}^{n} u_i x_i Q_i(2)\right]\left[\sum_{j=1}^{n} x_j^2 Q_j(2)\right] = \left[\sum_{i=1}^{n} u_i x_i Q_i(1)\right]\left[\sum_{j=1}^{n} x_j^2 Q_j(3)\right],$$

where $Q_i(d) := \prod_{(i\neq)m=1}^{n} (K_{\mathrm{M}} + C_{\mathrm{S},m})^d,\ i = 1, 2, \ldots, n$, and after some algebra,

$$\sum_{\substack{i,j=1\\(i\neq j)}}^{n} u_i x_i x_j^2 \left\{ s_i^2 s_j^2 \left(\prod_{(i,j\neq)k=1}^{n} s_k^4 \right) - s_i^3 s_j \left(\prod_{(i,j\neq)m=1}^{n} s_m^4 \right) \right\} = 0,$$

where $s_k := K_{\mathrm{M}} + C_{\mathrm{S},\mathrm{k}}$, $k = 1, 2, \ldots, n$. Finally, the polynomial of P of $\vartheta := K_{\mathrm{M}}$, as in (3.1), has to be solved. For each real-valued root of (3.1), i.e. for each possible estimate $\hat{K}_{\mathrm{M}}$, the corresponding $\hat{V}_{\max}$ estimate is then obtained through (18). Finally, among the real valued (of the total 4n – 5) candidate estimates, we choose as the NLLS estimate $\hat{\boldsymbol{\theta}} = (\hat{V}_{\max}, \hat{K}_{\mathrm{M}})$, the one which minimizes the sum of squares S.

Comparing the Linear Least Squares (LLS) estimations (2) and the analytic NLLS estimation of Theorem 3.1, the (treated) Puromicin dataset was adopted, as in Bates

and Watts (1998, Table A1.3). Table 1 presents the evaluated LLS estimators of V_{max} and K_M, corresponding to the Eadie-Hofstee (EH), Hanes-Wolf (HW) and the Lineweaver-Burk (LB) linearizations, together with the NLLS estimators $\hat{V}_{max}$ and $\hat{K}_M$ (obtained by applying Theorem 3.1), as well as their corresponding errors. It is clear that the NLLS estimation provides a "better" estimate than the LLS ones, in terms of the corresponding R_2 coefficients, *sse* (sum of squared errors) and *mare* (mean absolute squared errors). Notice that, the LLS (HW) estimated model approximates better the NLLS model, as adopts the least mean squared error among the other LLS models.

Method	$\hat{V}_{max}$	$\hat{K}_M$	R_2	*sse*	*mare* (%)
LLS (EH)	193.867711	0.043524	0.9282	184.69	10.74
LLS (HW)	216.216899	0.067915	0.9603	102.05	6.94
LLS (LB)	195.802709	0.048407	0.9378	160.05	9.19
NLLS	212.683743	0.064121	0.9613	99.62	6.99

Table 1. Comparison between the LLS and the analytic NLLS estimation.

Notice that, as the degree $(4n - 5)$ of the polynomial P in Theorem 3.1 is odd, there is always at least one (real-valued) estimator $\hat{\theta}$ for the M-M model. Moreover, as more than one estimators may exist, then more than one corresponding average-per-observation information matrices are possible. Therefore, the design might not be considered as unique. The criterion we suggest (the minimum *sse*) actually offers the design with the minimum variance, i.e. it corresponds to the D-optimal one.

REFERENCES

Bates, D.M., and Watts, D.G. (1998). *Nonlinear-Regression Analysis and its Applications*. Edinburgh: Oliver & Boyd.

Currie, D.J. (1982). "Estimating Michaelis-Menten Parameters: Bias, Variance and Experimental Design". *Biometrics*, 38, 907-919.

Dowd, J.E., and Riggs, D.S. (1965). "A Comparison of Estimates of Michaelis-Menten Kinetic Constants from various Linear Transformations". *J. of Biological Chemistry*, 240, 863-869.

Endrenyi, L., and Chan, F.Y. (1981). "Optimal Design of Experiments for the Estimation of Precise Hyperbolic Kinetic and Binding Parameters". *J. Theor. Biol.*, 90, 241-263.

Hill, P.D.H (1980). "D-optimal Designs for Partially Nonlinear Regression Models". *Technometrics*, 22, 275-276.

Kitsos, C.P. (2001). "Design Aspects for the Michaelis-Menten Model". *Biometrical Letters*, 38, 53-66.

Kitsos, C.P. (2012). *Cancer Bioassays: A Statistical Approach*. Lampert Acad. Pub.

Michaelis, L., and Menten, M.L. (1913). "Kinetics for Intertase Action". *Biochemische Zeiturg*, 49, 333-369.

54. SOLVING THE PORTFOLIO OPTIMIZATION PROBLEM UNDER REALISTIC CONSTRAINTS

R. Kizys
University of Portsmouth, Department of Economics and Finance, UK,
renatas.kizys@port.ac.uk

Á.A. Juan and A. Linares
IN3 – Open University of Catalonia, Department
of Computer Science, Spain,
{ajuanp, alinaresz}@uoc.edu

B. Sawik
AGH University of Science & Technology, Faculty of Management, Department of
Applied Computer Science, Poland,
b_sawik@cal.berkeley.edu

ABSTRACT

A hybrid algorithm for realistic portfolio optimization is introduced in this paper. Our matheuristic algorithm combines quadratic programming with a metaheuristic framework in order to solve complex variants of the mean-variance portfolio optimization problem. Thus, it is able to quickly generate near-optimal asset-allocation solutions —in terms of minimum risk for a given expected return rate— under realistic conditions. These realistic conditions include the well-known cardinality, preassignment, and quantity constraints. The algorithm proceeds in two steps. First, a feasible initial solution is constructed by allocating portfolio weights according to the individual return rate. Secondly, an iterated local search process gradually improves the initial solution throughout an iterative combination of a perturbation stage and a local search stage, which includes quadratic programming optimization. According to the experimental results obtained, our approach is competitive

when compared against existing state-of-the- art approaches, both in terms of the quality of the best solution generated as well as in terms of the computational times required to obtain it.

1. INTRODUCTION

Mean-variance optimization has become the workhorse model for portfolio selection since Markowitz (1952). An important assumption underlying this model is that investors are concerned about both the expected returns from their investment and the risk from that investment, where risk is defined as the variance of future returns. By optimally allocating appropriate weights to imperfectly correlated risky assets, investors can reduce the variance of future portfolio returns and thus diversify their investment. An optimal allocation of assets produces the so-called unconstrained efficient frontier (UEF), which consists of a 2D plot representation of portfolio configurations with a minimum degree of risk (horizontal axis) for each level of expected return (vertical axis) (Figure 1).

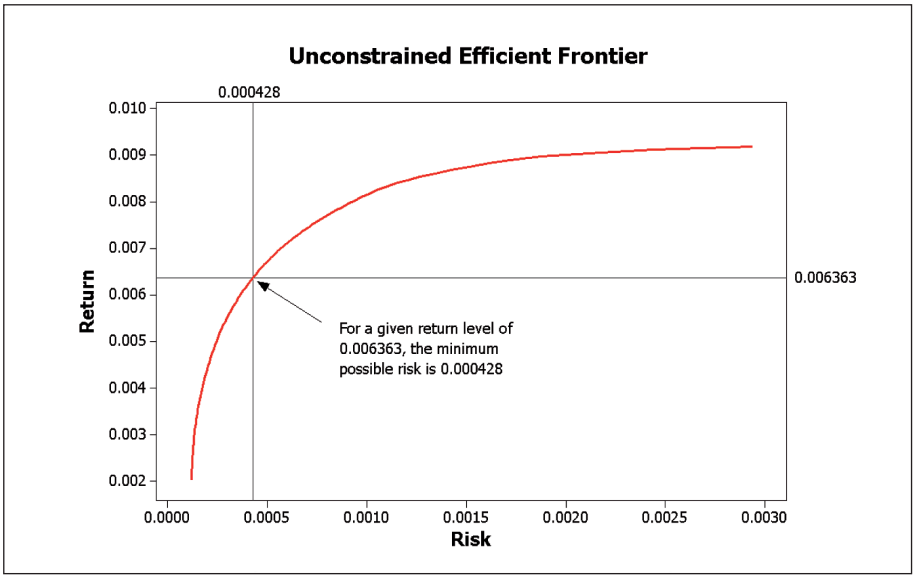


Figure 1. 2D plot representing the UEF.

Since the mean-variance portfolio formulation, an enormous amount of papers have been published extending or modifying the basic model in three directions (Sawik, 2013). The first path goes to simplification of the amount or the type of input data. The second path focuses on the introduction of an alternative measure of risk. The third path involves incorporation of the additional criteria and constraints. Traditional optimization methods include, among others, linear programming (Mansini *et al.*, 2014) and also quadratic programming (Sawik, 2012). Unfortunately, the complexity of the mean-variance model for portfolio optimization increases dramatically upon adding additional constraints that are needed to provide a realistic representation of the investor's portfolio choice, which reduces the efficiency of exact methods in other than small-size instances. Metaheuristic approaches have been proposed in the literature as a way to surpass these limitations (Maringer, 2005). This paper focuses on the single-period version of the so-called constrained mean-variance portfolio optimization problem. We will assume that all assets are risky assets (stocks). However, the solving approach introduced here could be also used for similar optimization problems regarding commodities, futures, options and swaps. The objective function is the portfolio variance, which is considered to be the measure of risk on investment in the portfolio. It is obtained as a weighted average of return co-variances between each pair of assets included in the portfolio. The decision problem involves minimizing the portfolio variance while ensuring a return level not inferior to the one required by the investor. Portfolio weights add up to one and are constrained to take on non-negative values only. Real-world portfolio selection problems may involve additional constraints. The adoption of an initial solution that is based on a well-chosen criterion makes our solver more flexible when solving tight instances. Also, the rate of return on an individual asset is used as a main criterion to construct the initial solution. It is worth noting that this criterion provides the best possible solution in terms of guarantying feasibility of the requested portfolio return —i.e., if this initial solution is not feasible, then the problem has no feasible solution. By contrast, in most of the existing approaches, an initial solution is randomly drawn and, hence, the feasibility of these initial solutions cannot be guaranteed. Finally, our solver features several efficient sub-procedures that seek to avoid getting trapped into a local optimum. For a comprehensive overview of the formulations of the portfolio selection problem, see Di Tollo and Rolli (2008).

2. FORMAL DESCRIPTION OF THE PROBLEM

A set of n assets is given by the market, $A = \{a_1, a_2, ..., a_n\}$, where:

- $\forall i \in \{1, 2, ..., n\}$, ai has a known expected return, $r_i \geq 0$.
- $\forall i, j \in \{1, 2, ..., n\}$, the pair (a_i, a_j) has a known expected risk index, $\sigma_{ij} = \sigma_{ji} \in R$.
- Expected return constraint: a user-provided value, $R > 0$, represents the minimum expected return from the investment.
- A portfolio is a vector $X = (x_1, x_2, ..., x_n)$ such that each x_i represents the fraction of the total budget invested in asset a_i, i.e., $0 \leq x_i \leq 1$ and $\sum_{i=1}^{n} x_i = 1$.
- $\forall i \in \{1, 2, ..., n\}$, $z_i = 1$ if $x_i > 0$ (i.e., a_i in portfolio) and $z_i = 0$ otherwise.
- Cardinality constraint: the number of assets in the portfolio, $\sum_{i=1}^{n} z_i$, is bounded by user-defined values, $k_{\min}$ and $k_{\max}$.
- Quantity constraints: for each asset a_i, its associated quantity in the portfolio, x_i, is bounded by user-defined values, ε_i and δ_i.
- Pre-assignment constraints: the user can pre-select certain assets to be included in the portfolio. $\forall i \in \{1, 2, ..., n\}$, $p_i = 1$ if a_i is pre-assigned in portfolio (i.e., $x_i > 0$) and $p_i = 0$ otherwise

Then, the problem consist in selecting the optimal combination of fractions of each asset, so that the overall variance (risk) is minimized while satisfying all the aforementioned constraints. More formally:

$$\min f(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j$$

subject to:

$$\sum_{i=1}^{n} r_i x_i \geq R \tag{1}$$

$$\sum_{i=1}^{n} x_i = 1 \tag{2}$$

$$k_{\min} \leq \sum_{i=1}^{n} z_i \leq k_{\max} \tag{3}$$

$$0 \leq x_i \leq 1, \ \forall i \in \{1,2,...,n\} \tag{4}$$

$$\varepsilon_i z_i \leq x_i \leq \delta_i z_i, \ \forall i \in \{1,2,...,n\} \tag{5}$$

$$x_i \leq z_i \leq p_i, \ \forall i \in \{1,2,...,n\} \tag{6}$$

$$z_i \in \{0,1\}, \ \forall i \in \{1,2,...,n\} \tag{7}$$

3. SOLVING APPROACH

The proposed matheuristic makes use of an Iterated Local Search (ILS) framework (Lourenço *et al.* 2003). ILS is a conceptually simple yet powerful metaheuristic that has proven to be very efficient in solving complex combinatorial optimization problems. The underlying idea behind ILS is to narrow the search for candidate local optimal solutions returned by some embedded algorithm, typically a local search heuristic. At some stages of the ILS, a quadratic programming optimizer is called as a local search in order to improve the investment levels assigned to each of the assets in the current portfolio. Figure 2 provides a high-level description of our approach (the low-level details are not included here due to space limitations).

```
01 Procedure Matheuristic(inputs)
02 baseSol <- generateInitialSolution % generate a feasible solution
03 bestSol <- baseSol
04 while {elapsedTime <= maxTime} do
% modify portfolio configuration without losing feasibility
05   newSol <- perturbate(baseSol)
% improve assets levels via quadratic programming
06   newSol <- localSearch(newSol)
07   if {risk(newSol) < risk(baseSol)} then
08     baseSol <- newSol
09     if {risk(newSol) < risk(bestSol)} then
10       bestSol <- newSol
11     end if
12   end if
13 % (optional) include an acceptance criterion
14 end while
15 return bestSol
16 end procedure
```

Figure 2. Overview of our matheuristic approach.

4. COMPUTATIONAL EXPERIMENTS

The algorithm has been implemented as a Java application. A standard personal computer, Intel Core i5 CPU at 3.2 GHz and 4 GB RAM with Linux Ubuntu, was used to perform all tests. In this research, we experiment with sets of stock market data already used in previous studies by Moral-Escudero *et al*. (2006) and Di Gaspero *et al*. (2011). The data set comprises constituents of five stock market indices, Hang Seng (Hong Kong), DAX 100 (Germany), FTSE 100 (United Kingdom), S&P 100 (United States) and NIKKEI 225 (Japan). Figure 3 shows a comparative among different approaches, where gap refers to the gap with respect to the unconstrained problem – lowerbound for the optimal solution of the constrained problem. Notice that our algorithm is able to outperform the other approaches in 4 out of the 5 sets.

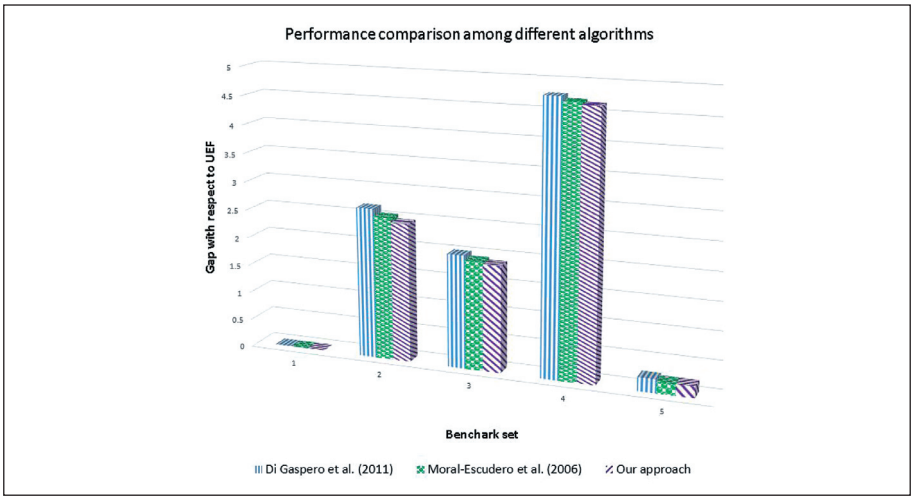


Figure 3. Visual comparison among different approaches.

5. CONCLUSION

We have proposed hybrid algorithm, combining a metaheuristic framework with quadratic programming, to efficiently deal with the constrained mean-variance optimization problem. The specific constraints we use in our tests include the

cardinality and quantity constraints. The ensuing complexity of the problem rules out closed-form solutions and conventional optimization methods. Specifically, the cardinality and quantity constraints render the problem computationally expensive to solve real-world instances.

Therefore, the use of metaheuristics —that handle such problems more flexibly and efficiently— is needed. According to the computational experiments, our solver outperforms —in terms of average percentage loss— some of the most recent approaches in the literature.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (grant TRA2013-48180-C3-3-P) and FEDER. Likewise we want to acknowledge the support received by the Department of Universities, Research & Information Society of the Catalan Government (Grant 2014-CTP-00001) and by the NCN grant (6459/B/T02/2011/40).

REFERENCES

Di Gaspero, L., Di Tollo, G., Roli, A., and Schaerf, A. (2011). "Hybrid Metaheuristics for Constrained Portfolio Selection Problems". *Quantitative Finance* 11 (10), 1473-1487.

Di Tollo, G., and Roli, A. (2008). "Metaheuristics for the Portfolio Selection Problem". *International Journal of Operations Research*, 5 (1), 13–35.

Lourenço, H.R., Martin, O.C., and Stützle, T. (2003). "Iterated Local Search". *Handbook of Metaheuristics*, 321-353. Kluwer. Norwell, MA.

Mansini, R., Ogryczak, W., and Speranza, M.G. (2014). "Twenty Years of Linear Programming Based Portfolio Optimization". *European Journal of Operational Research*, 234 (2), 518-535.

Maringer, D. (2005). *Portfolio Management with Heuristic Optimization*. Springer.

Markowitz, H.M. (1952). "Portfolio ". *Journal of Finance*, 7 (1), 77-91.

Moral-Escudero, R., Ruiz-Torrubiano, R., and Suárez, A. (2006). "Selection of Optimal Investment Portfolios with Cardinality Constraints". *IEEE Congress on Evolutionary Computation*, 2382-2388.

Sawik, B. (2012). "Bi-Criteria Portfolio Optimization Models with Percentile and Symmetric Risk Measures by Mathematical Programming". *Przeglad Elektrotechniczny*, 88 (10B), 176-180.

Sawik, B. (2013). "Survey of Multi-Objective Portfolio Optimization by Linear and Mixed Integer Programming". *Applications of Management Science*, 16, 55-79.

55. EFFECTIVE RISK ASSESSMENT REVIEW (FOR THE POWER UTILITIES)

Leo Pui Yong
Tenaga Nasional Berhad, Risk Management Department, Malaysia,
leopy@tnb.com.my

ABSTRACT

Engineers often view risk managment as an academical exercises. As engineers of power utilities, it is always their concern on uninterrupted supply, no equipment failure and many others. This paper introduced a step-by-step approach on effective risk assessment review, using some business entities related to power industry as example. An effective risk assement review helps to further enhance risk management system in respective business entity.

1. INTRODUCTION

According to the Committee of Sponsoring Organizations of the Treadway Commission (COSO), Enterprise Risk Management (ERM) is defined as a process, affected by an entity's board of directors management and other personnel, applied in strategy setting and across the enterprise, designed to identify potential events to be within risk appetite, to provide reasonable assurance regarding the achievement of entity objectives.

Based on the definition, the determining factors for successful ERM implementation are derived as follows:

a. Tone of the top.

b. Approved ERM framework.

c. Effective communication across the enterprise.

d. Risks identified in alignment with the company's business objectives.

e. Effective and efficient controls and mitigation plans.

In order to carry out effective risk assessment review for a business entity, the above factors must be assessed.

2. EFFECTIVE RISK ASSESSMENT REVIEW

The objectives of the Risk Assessment Review should be identified as follows:

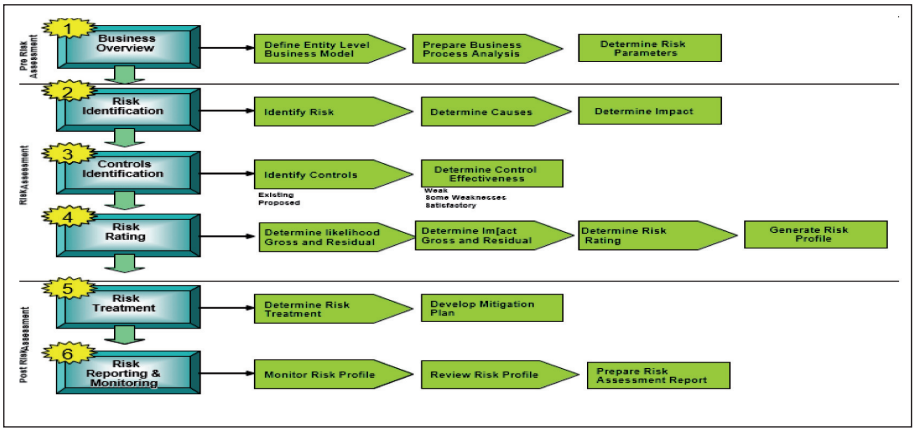
2.1. To establish the existence and adequacy of "tone of the top" for risk management

In Malaysia, it is widely promoted for the Management to be responsible for the governance of risk. According to the Malaysian Code of Corporate Governance 2012 (MCCG 2012), the Board for public listed companies is responsible for satisfying itself that the Management has developed and implemented a sound system of risk management. The "hard control" clearly exist to promote the importance of risk management in an organisation. The "tone of the top" will be more apparent when the "soft control" is added within the top management. The top management must set the tone to the employees that risk management is important even in a day to day operation. All employees must think of risk before any action or decision is being undertaken. Therefore, it is crucial to establish the "tone of the top" for risk management, i.e. top management believe in risk management and these voices is clearly communicated to every staff in the organisation.

2.2. To establish the existence and adequacy of risk management framework

Adequate risk management framework should be approved or endorsed for implementation across the organisation. The risk management framework can be established in alignment with the existing standards, such as: ISO 31000 and COSO ERM

Framework. A typical risk management framework consists of the following steps" Business Overview, Risk Identification, Controls Identification, Risk Rating, Risk Treatment, Risk reporting and monitoring as shown below:



The risk management framework should also incorporate guidelines for the followings:

- Risk Parameters - Determining risk parameters or a category scale of measurement for estimating both the likelihood of a risk occurring and the impact of that risk if it does occur. In this case, the criteria should be stated to represent the Group's risk appetite. Example, for a specific power plant (e.g. 3 X 700MW), the risk appetite might be Unplanned Outage Rate (UOR); however, the Group's risk appetite, (i.e. Grid System Operator) might be different. As a Grid System Operator, the concern maybe not having unplanned outages all the 3 units of 700MW at the same time. This is very important, in fact may be the determining factor on how much controls or mitigation plans to be put in as risk treatment.

- Likelihood – The criteria for different category of likelihood should be stated for easy references and decision making. For example:

Category of Likelihood	Criteria
Rare	Event may occur only in exceptional circumstances, e.g., approximately below 5% chance of occurring in the next 12 months
Unlikely	The event could occur at some time, e.g. approximately below 25% but above 5% chance of occurring in the next 12 months.
Moderate	The event might occur at some time, e.g. approximately below 50% but above 25% chance of occurring in the next 12 months.
Likely	The event will probably occur in most circumstances, e.g. approximately below 95% but above 50% chance of occurring in the next 12 months.
Almost Certain	The event is expected to occur in most circumstances, e.g. approximately above 95% chance of occurring in the next 12 months.

The above criteria can be more quantitative when they are linked to historical statistic. For example, one of the risk in a power utility company might be "Overhead Lines Tripping with Load Loss". The criteria for the likelihood can be related to the actual tripping data if a monitoring mechanism is put in place and root cause analysis is being carried out as follows:

No	Tripping Incident	Date	Root Cause	Load Loss
1.	132kV Overhead Line from Substation A - B	XX Sep 2010	Lightning	XX MW
2.	275kV Overhead Line from Substation C - D	XX October 2010	Protection Mal operation	XX MW
3.	132kV Overhead Line from Substation E - F	XX October 2010	Aging Equipment	XX MW
Total Overhead Line Trippings		XX incident		XX MW

The above statistic can be relate to the criteria of likelihood for a basis that is quantitative.

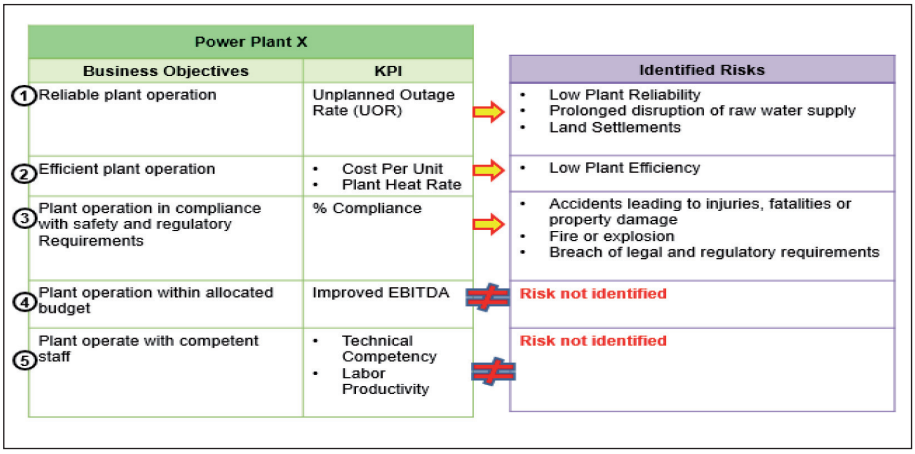
2.3. To establish the existence, adequacy and effectiveness of controls to ensure compliance of the above risk management framework

After establishing the adequacy and existence of a risk management framework, it is important to establish if the approved risk management framework is complied with. In this case, the Risk Assessment Review must first confirm if there are

controls available to ensure compliance of the above risk management framework, i.e. procedure for implementation, monitoring process, and reporting mechanism. Subsequently, random tests (via sampling) can be conducted to assess if the details stated in the risk management framework is complied with.

2.4. To establish if risks identified are aligned with the business objectives

Risks identified should be aligned with the business objectives. A simple assessment would be "mapping" identified risks to the business objectives. One of the example of the "mapping" exercise conducted for a power plant is as shown below.



Power Plant X has typical business objectives, i.e. to operate the power plant in a manner that is reliable, efficient, in compliance with safety and regulatory requirements. At the same time, the operation of the power plant must be kept within allocated budget and the operation must be carried out by competent staff. In this case, the Key Performance Indicators are derived from the business objectives. As risk management should be embedded in the day-to-day business operation, identified risks should be aligned with the business objectives and not otherwise.

2.5. To establish the existence, adequacy and effectiveness of controls and mitigation plans for the identified risks

Adequate controls should be available to mitigate the identified risks. Assessment of each control or mitigation plan must be conducted effectively. One of the case study for Power Plant Y, whereby a total plant shut down (i.e. 3 X 700MW) has caused load shedding in the country.

Power Plant Y has identified one of the risks as low plant reliability. Some of the root causes for the risk of low plant reliability include: ineffective maintenance, critical equipment failure and design failure. One of the control introduced is Reliability Centered Maintenance (RCM). RCM is an engineering framework that enables the definition of a complete maintenance regime. It regards maintenance as the means to maintain the functions a user may require of machinery in a defined operating context. As a discipline it enables machinery stakeholders to monitor, assess, predict and generally understand the working of their physical assets. In summary, the key concept is "The task shall reduce the probability of the failure mode to a level that is tolerable to the owner or user of the asset," implying that the RCM team will have to judge whether an acceptable level of risk is achieved by implementing the selected maintenance tasks.

In the case of Power Plant Y, there are total of 200 systems which requires RCM to be conducted. Management of Power Plant Y decided to prioritize the systems for detailed RCM based on the historical failure data. RCM was carried out for 38 out of 200 systems in Power Plant Y prior to the incident whereby all 3 units of 700MW was on forced outage due to one of the other system's failure. Subsequent to the incident, review on the risk profiles disclosed if the management of Power Plant Y had identified earlier that they cannot afford to lose all the 3 units of 700MW at the same time, the risk profiles would have been different. In this case, priority will be given to identify all the systems (from the 200 systems) that can caused simultaneous forced outage of all units.

As a result of risk assessment review, it is always a good practice to develop a simple rating mechanism to inform the respective parties the risk maturity level. An example of rating mechanism is as shown in the next page.

AREA	RATING			
	BEGINNER	INTERMEDIATE	PROGRESSIVE	MATURE
Communication	Unaware of risk management concept, EWRM policy and GWRA at all levels.	Top management are aware of risk management concept, EWRM Policy and GWRA but do not disseminate the information to the staff.	Top management are aware of risk management concept, EWRM Policy and GWRA and disseminate the information to the staff.	Top management are aware of risk management concept, EWRM Policy and GWRA and disseminate the information to the staff. Additionally, ensure comprehensive understanding and application. "Think of risk first at all levels"
Risk Management Structure	Risk management structure not available.	Risk management structure is available but no clear segregation of roles and responsibilities.	Risk management structure (i.e. risk owner, risk manager, risk coordinator and etc.), is available but not effectively implemented.	Risk management structure (i.e. risk owner, risk manager, risk coordinator and etc.), is available and effectively implemented.
Non-compliance to EWRM policy and GWRA	>50 non-compliances observed during risk assessment review.	30-49 non-compliances observed during risk assessment review.	1-29 non-compliances observed during risk assessment review.	No non-compliance observed during risk assessment review
Alignment of identified risks to business objectives	Risks are identified without considering the TNB strategic risks, 5 KRAs and KPIs.	Risk registers are not aligned to TNB strategic risks, 5 KRAs and KPIs.	Risk registers are partially aligned to TNB strategic risks, 5 KRAs and KPIs.	All risk registers aligned to TNB strategic risks, 5 KRAs and KPIs.
Control adequacy & effectiveness	Controls do not adequately address the identified causes.	Controls partially address identified causes.	Controls adequately address the identified causes but not effectively implemented.	Controls adequately and effectively address the identified causes.

3. CONCLUSION

It is important to obtain buy in at all level that the results of the risk assessment review not only representing the risk maturity level of each business entity, but also useful resources to review and update the business entity's risk management system.

ACKNOWLEDGEMENTS

Special thanks to Tenaga Nasional Berhad, the main sponsor of the paper research. Special thanks to the Board Risk Committee of Tenaga Nasional Berhad, CEO and CRO of Tenaga Nasional Berhad for the effort in adopting good risk management system in the company.

REFERENCES

Moeller, R.R. (2011). *COSO Enterprise Risk Management: Establishing Effective Governance, Risk and Compliance Process*. Wiley Corporate F&A.

Moubray, J. (1997). *Reliability-Centered Maintenance* (second edition).

56. SILVER SUPPLY RISK ASSESSMENT FOR SUSTAINABLE. PHOTOVOLTAIC INSTALLATION

W. Li
Akita University, Graduate School of Engineering and Resource Science, Japan, gentlelwh@163.com

T. Adachi
Akita University, Faculty of International Resource Science, Department of Resource Policy and Management, Japan, adachi.t@gipc.akita-u.ac.jp

ABSTRACT

Silver is one of essential raw materials for Crystalline Silicon (C-Si) solar cells and is used as electrode. C-Si type cell holds roughly 90% of the photovoltaic (PV) market presently. Therefore, it arouses general concern whether silver is able to secure PV installation in the long term or not. Main aim of this research is to measure the supply availability of silver used for PV in the future. Some regression models along with four scenarios are applied to predict the potential of PV installation. Primary silver supply is estimated by Hubbert Curve model. Silver demand for PV is obtained by taking material substitution and conversion efficiency improvement into account. Furthermore, five indicators are introduced to assess silver supply risk in present. For one of the risks, we develop a quantitative country risk model for silver mine investment by combining linear-log regression and optimization method. The result shows that silver demand for PV is more sensitive to substitution than efficiency improvement. And, supply shortage may occur around 2030 based on government planning. The present supply risk is tend to be moderate through five years observation. In terms of investment, Australia, America, Chile, Canada, Poland were found to be more attractive because of lower risk.

1. INTRODUCTION

Renewable energy will take place of conventional sources along with the depletion of fossil energy and the deteriorating environment. PV solar cell is one of promising alterative renewable energy which can generate electricity without resource limitation. Moreover, it quickly expanded worldwide in the last decade from less than 2GW installed electricity capacity in 2000 to almost 140GW by 2013 by EPIA (2014). In 2011, 2.2% of world silver demand was consumed in the electrode of solar panels (Leena and Andrea, 2014). Our first aim focuses on how much silver will be required in the next 35 years, followed by comprehensive silver supply risk analysis by different indicators.

Leena and Andrea (2014) applied a method to estimate material cumulative upper limit and annual growth constraints based on given technologies. In our research, dynamic approach is applied to estimate silver availability. In raw material supply risk analysis, Benjamin and Christoph (2013) induced an overview of 15 supply risk evaluation methods where 21 indicators are compared, while 5 indicators are selected and aggregated to distinguish the level of silver supply risk in this study. Regardless of numerous qualitative country risk assessment, country risk analysis for mining investment is developed by combination of financial, political and resource potential factors.

2. MODEL AND METHODOLOGY

Total installed generation capacity of electricity (EIA) has a linear relation with GDP of the world in current US dollars World Bank. Continuous growth of world population (World Bank) will eventually depend on resource availability. We assume growth rate of population will have exponential changes in the beginning of this century and the slope of the growth rate become lower and lower in the middle of century, and asymptote to zero finally. Thus, an S-shaped logistic regression by time is applied to world population prediction. Adding a function of population and GDP, generation capacity of electricity in the future is forecasted in our model, shown in equation 1, where:

- $Popsat$ represents population at saturation level.

- Et represents world total generation capacity of electricity in year t.

- Others are constants.

$$E_t = m \times \left(\frac{Pop_{sat}}{1 + e^{a+bxt}} \right)^c + n \quad (1)$$

Four scenarios shown in table 1 is adopted to predict the world installed photovoltaic capacity. In the medium scenario, 20% of world total installed electricity capacity will be generated from PV by the end of 2050 was assumed at exponential growth trend. This forecast is close to the result of IEA, Technology Roadmaps which predicted PV capacity will hold 3,000GW by 2050. The policy (4+) represents policy based scenario of four countries (China IEA-PVPS, 2012; Japan IEA, 2009, 2012; EU EPIA, 2010; USA EIA, 2012), since their installation of PV covers 90% of the total installation recently, according to EPIA (2014).

Silver primary supply forecast is based on a linear regression between ratio of annual production USGS and cumulative production, and cumulative production itself. Related to silver demand for PV, improvement of conversion efficiency of PV cell (2008, 16%-2050, 40%) and material substitution to copper (2012, 8.22g/m^2-2023, 2.05g/m^2) predicted by IEA (2010), ITRPV (2012) is applied with linear growth. Silver usage per unit of PV cell will decrease to 0.822g/m2 by 2050 to represent upper limit substitution according to Leena and Andrea (2014). Silver supply risk analysis are developed by aggregated five indicators which is shown in table 2.

Country risk modeling for silver investment contains financial risk, political risk and reciprocal of resource potential. To measure these risks, data of 17 silver producing countries which cover 95% of total production and more than 90% of reserve (USGS) from 2010 to 2012 are observed. Financial risk is represented by weighted sum of 8 variables (WB, IMF) shown in table 3. In case some variables contain negative values, special logarithm $\ln\left(x + \sqrt{x^2 + 1}\right)$ is adopted to transform the data. To obtain coefficients of these variables against financial risk level, we induce an optimization method to minimum error between Standard &and Poor's sovereign ranking and our calculated results. Political risk is also obtained in the same optimization for

different other 6 variables (WB, IMF) shown in table 4. Comparative ranking source is world governance indicator from World Bank in this case. Resource potential for specific countries are considered as ratio of its reserves share and production share of world total. All three results are adjusted to 0 to 5 integer and then aggregated on radar chart to make comparison easier.

Scenario	Share	EPV(GW) (2050)
High	40%	6168
Medium	20%	3084
Low	10%	1542
Policy(4+)	19%	2933

Table 1. Four forecast scenario for PV installation.

Risk Indicators	Definition	High risky 5.0-4.1	Risky 4.0-3.1	Moderate 3.0-2.1	Low risk 2.0-1.1	Adequate 1.0-0.1
Depletion Time (DT) (USGS 2009-2013)	R/T Reserve divided by annual mining production	0-10 years	0-10 years	0-10 years	0-10 years	0-10 years
Price Volatility (PV) (London Fix)	$\sqrt{\frac{\sum(k-\bar{n})^2}{n-1}}$ Standard deviation of price change ratio	200%-51%	50%-31%	30%-16%	15%-6%	5%-0%
Recycling Ratio (RR) (WWS 2014)	P_S/P Annual production from old scrap divided by production of mining and old scrap	0%-5%	6%-15%	16%-30%	31%-50%	Over 51%
By production ratio (BP) (WWS 2014)	P_B/P By-production of base metal divided by production	100%-61%	60%-41%	40%-31%	30%-11%	Less than 10%
Concentration Ratio (CR) (WWS 2014)	HHI Sum of square of production share by country	10,000-5,001	5,000-2,501	2,500-1,501	1,500-101	100-0

Table 2. Supply risk indicators definition and aggregation standard.

X1	X2	X3	X4	X5	X6	X7	X8
Debt/ GDP	Reserve/ Debt	Debt/ Export	GDP per capita	GDP growth rate	Inflation	Account balance/ GDP	Government revenue/GDP

Table 3. Variables for financial risk modeling.

X1	X2	X3	X4	X5	X6
Unemployment	Rural population	GDP per capita	Improved water source	Inflation	GDP growth rate

Table 4. Variables for political risk modeling.

3. RESULTS

The world primary silver production in 2012 was 25,500 ton. The production is expected to reach 31,100 ton in 2030, and 43,000 ton in 2050 under our estimate, shown in figure 1. The average growth rate is 1.7% annually and is varied from 1.8% to 1.5%.

For policy based scenario, the cumulative silver demand for PV is likely to be 30,000 ton by effective substitution, which is around 1% of total primary silver supply. However, without substitution, silver demand will increase significantly, up to 3.5-5.5% of the total by 2050. Therefore, improvement conversion efficiency is not enough effective to mitigate supply restriction of silver. For annual demand, fluctuations of policy (4+) scenario in 2020 and 2030 in figure 3 are caused by the uneven growth rate of PV installation plan of governments. It implies another important prospect that supply restriction might occur around 2030 due to high growth rate for the first two decades.

For silver supply risk assessment, life time of discovered reserves recovered after 2009, because of increasing exploration in gold, copper, lead and zine deposits. These base metal mines also produce silver as a by-production, and about 70% of silver are produced in these base metal mines. As silver price increasing continuously after 2003, average price volatility of silver also went high from 20% up to 35%, which indicates higher risk. Price volatility is the most sensitive indicator compared to the other four indicators. Silver recycling ratio fluctuated around 25% before 2000,

after that, it declined to below 25% with a negative trend because digital cameras have been increasingly replacing film cameras which consume silver for film and developer. Concentration ratio of silver market measured by HHI (Herfindahl-Hirschman Index) is lower than 1,500, which indicates silver market is competitive one (low risk). From the results for five years shown in the figure 4, silver supply risk of silver has been slightly reducing and the risk level tends to be moderate. Besides that, trend of by-production ratio stays constant in high level, which means the supply flexibility is low for silver.

Finally, the country risk for mining investment is conducted to look at new market opportunity. Results for 2012 is shown in figure 5. Australia, America, Chile, Canada and, Poland can be categorized as more attractive investment environment because of low risk. Among them, Australia has lowest financial risk, and Sweden has lowest political risk. While Chile and Poland have highest resource potential, followed by Australia. Morocco, Peru, China, Kazakhstan, Russia, Turkey, and Indonesia are moderate risk investment countries, in which Peru and China holds higher resource potential. Mexico, Argentina, Bolivia and India are categorized into high risky countries for investment, where Mexico actually has lower financial risk and political risk but less resource potential.

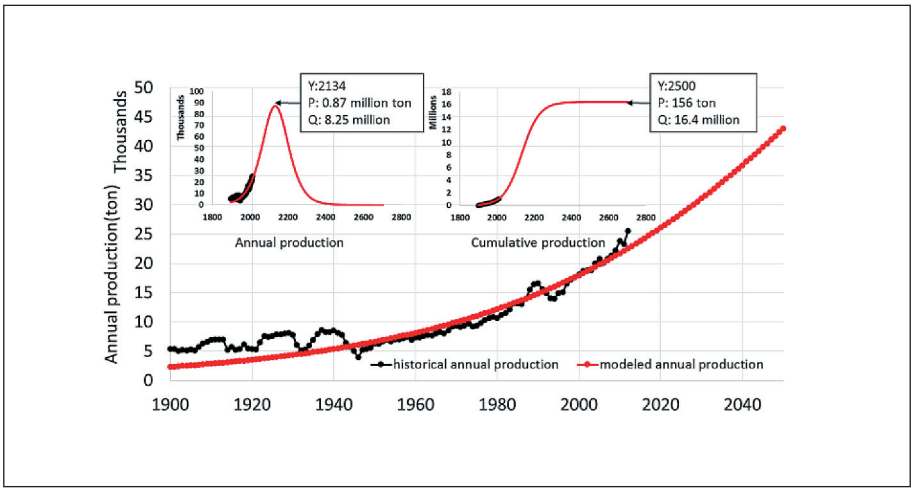


Figure 1. Primary silver supply forecast.

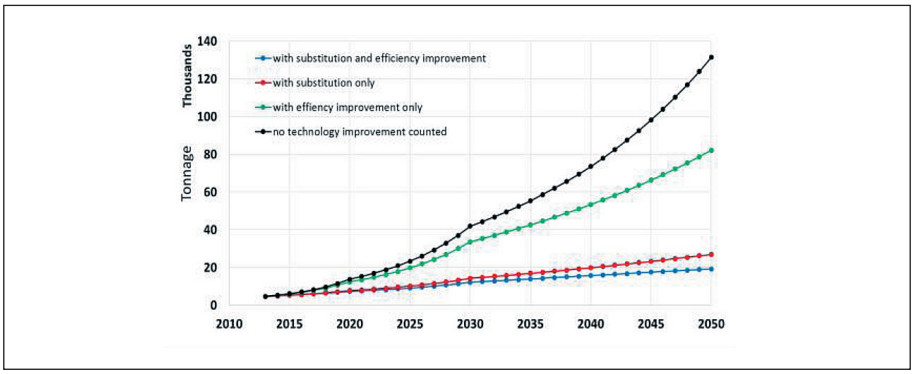


Figure 2. Cumulative silver demand forecast in the policy based scenario.

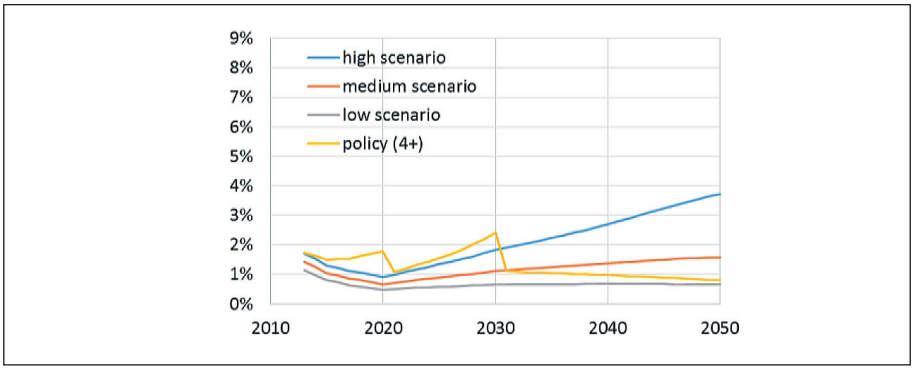


Figure 3. Silver demand in total primary supply forecast with substitution and conversion efficiency improvement.

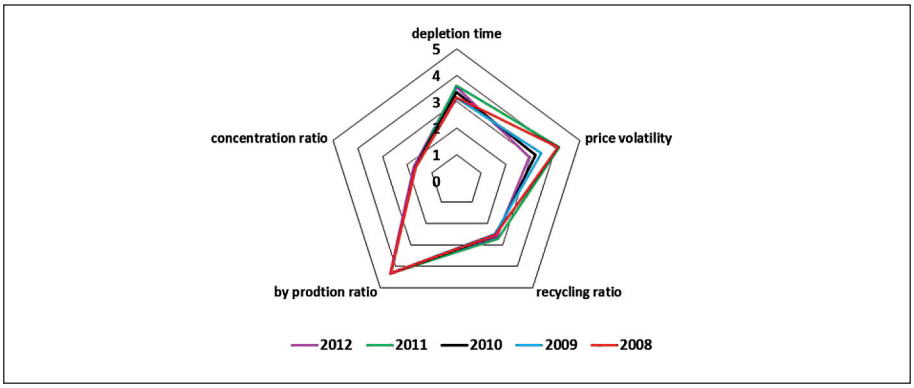


Figure 4. Silver supply risk assessment of 2008-2012.

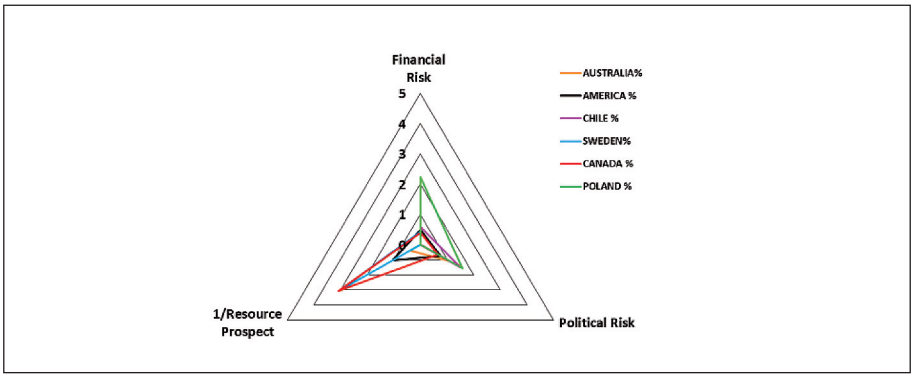


Figure 5. Country risk assessment of 2012 (low risk countries).

4. CONCLUSION

In short, looking at the balance of silver supply and demand, supply restriction might occur in the future because of high growth rate of PV installation. By risk assessments of silver supply, we sort its supply structure and find several weak points. Then, country risk analysis is beneficial to find future supply opportunity. After more detailed 3541 arrangement and comparison with other metals, the analysis in this study can contribute to suggestion for effective policy making and investment.

REFERENCES

Achzet, B., and Helbig, Ch. (2013). "How to Evaluate Raw Material Supply Risks. An Overview". *Journal of Resource Policy*, 38, 435-447.

EIA (2012). *SunShot Vision Study*.

EPIA (2010). *Solar Generation 6 Executive Summary.*

EPIA (2014). *Global Market Outlook for Photovoltaics 2014-2018*.

Grandell, L., and Thorenz, A. (2014). "Silver Supply Risk Analysis for the Solar Sector". *Journal of Renewable Energy*, 69 (2014), 157-165.

IEA (2009). *National Survey Report of PV Power Applications in Japan 2009*.

IEA (2012). *National Survey Report of PV Power Applications in Japan – 2012*. (Third version).

IEA (2010). *Technology Roadmaps Solar Photovoltaic Energy*.

IEA-PVPS (2012). *National Survey Report of PV Power Applications in China 2012*.

ITRPV (2013). *International Technology Roadmap for Photovoltaic 2012 Results*.

WWS (2014). *World Silver Survey 2014*.

57. NUEVA PROPUESTA DE ORDENACIÓN DE COMPAÑÍAS ASEGURADORAS

S. Linares-Mustarós
Universidad de Girona, Departamento de Empresa, España,
salvador.linares@udg.edu

J. M. Merigó-Lindahl
Universidad de Chile, Departamento de Control de Gestióny Sistemas de Información, Chile,
jmerigo@fen.uchile.cl

A. M. Gil-Lafuente
Universidad de Barcelona, Departamento de Economía y Organización de Empresas, España,
amgil@ub.edu

RESUMEN

El presente trabajo tiene como objetivo principal construir una metodología para establecer un *ranking* de compañías aseguradoras basado en el criterio de mejor adaptación a las necesidades demandantes de un posible cliente. Se ha escogido desarrollar una metodología sustentada en la nociones de "distancia" y "semidistancia" entre oraciones enunciativas cuyo valor de verdad se encuentra en el intervalo [0,1]. Un ejemplo numérico esclarece las notaciones y muestra el potencial del nuevo índice.

1. INTRODUCCIÓN

El trabajo ha sido elaborado aceptando el supuesto sugerido por Black (1937) y Zadeh (1965) de que es posible asignar a determinadas oraciones un cierto grado de

verdad y falsedad en el intervalo [0,1]. Dicha asunción permite modelizar problemas (Kaufmann y Gil Aluja, 1986; Gil-Lafuente, 1993) obteniendo soluciones más cercanas a nuestra forma de pensar que las que se obtienen al realizar modelizaciones basadas en la teoría de lógica clásica (Linares y Ferrer, 2014).

Esperamos mostrar una metodología de clasificación de compañías aseguradoras basada en la lógica borrosa.

Mediante el uso de la nueva metodología se podrá establecer un orden de prelación entre diversas compañías susceptibles de ofrecer un seguro específico, obteniendo un listado ordenado desde la empresa más adecuada para la demanda hasta la menos idónea.

Consiguientemente, el beneficio que aportaría la nueva metodología consistiría básicamente en la obtención de una información coherente sobre qué compañía conviene contratar teniendo en cuenta características que hasta hoy no se habían tenido en cuenta.

Entre los diversos instrumentos ofrecidos por las modernas técnicas operativas de gestión se ha escogido desarrollar la metodología sustentándonos en la noción de "distancia".

Dicho concepto expresa, en cierto modo, el grado de "alejamiento" existente entre dos objetos físicos o mentales. En nuestro caso, reflejará el alejamiento entre el perfil ideal de una empresa aseguradora que cumple exactamente unas cualidades de valoración numérica subjetiva estimables por un cliente y el perfil ideal estimado por dicho cliente de las empresas potenciales.

2. METODOLOGÍA DE CLASIFICACIÓN

Se define formalmente una distancia, como una aplicación sobre los reales entre dos elementos de un conjunto X, que cumple las siguientes condiciones:

1. $d(x, y) \geq 0 \quad \forall\, x, y \in X$

2. $d(x, y) = d(y, x) \quad \forall\, x, y \in X$

3. $d(x, x) = 0 \quad \forall\, x \in X$

4. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall\, x, y, z \in X$

5. Si existen dos elementos $x, y \in X$ tales que $d(x, y) = 0$, entonces $x = y$

En caso de no cumplirse la propiedad 5, se denomina a la aplicación "d" como "semidistancia".

A continuación vamos a presentar dos funciones distancia que serán la base de nuestro futuro proceso de *ranking* entre compañías aseguradoras.

Dado el conjunto $X = [0,1]^n = [0, 1] \text{ x } [0, 1] \text{ x } \ldots_{(n-3)} \text{ x } [0,1]$, se define la distancia de Hamming[1] entre dos elementos $x = (x_1, x_2, x_3, \ldots, x_n)$ con $x_i \in [0, 1]$ e $y = (y_1, y_2, y_3, \ldots, y_n)$ con $y_i \in [0, 1]$ como:

$$d_H(x,y) = \sum_{i=1}^{n} \left|x_i - y_i\right| = \left|x_1 - y_1\right| + \left|x_2 - y_2\right| + \left|x_3 - y_3\right| + \ldots + \left|x_n - y_n\right|$$

Dados dos elementos x e y de un conjunto $X = [0,1]^n$ y unos pesos sobre cada componente w_i, tales que $\sum_{i=1}^{n} w_i = 1$, se define la distancia de Hamming ponderada mediante la función:

$$d_{Hw}(x,y) = \sum_{i=1}^{n} w_i \left|x_i - y_i\right| = w_1 \left|x_1 - y_1\right| + w_2 \left|x_2 - y_2\right| + \ldots + w_n \left|x_n - y_n\right|$$

Entre las cualidades más apreciadas de las distancias de Hamming destaca su gran sencillez y facilidad de cálculo. Dichas cualidades hacen de la distancia de Hamming una herramienta muy operativa y fiable en procesos de ordenación.

[1] En otras áreas de conocimiento es habitual denominar a dicha distancia como "distancia de Manhattan", "distancia L1" o "distancia de Minkowsky con orden 1".

La práctica aconseja normalmente el empleo de la herramienta distancia de Hamming en el supuesto de que se acepte que para cada característica requerida, tanto si el nivel poseído por un candidato no llega al nivel ideal como si lo sobrepasa, resulta igualmente menos valorada.

Habitualmente, no siempre se está de acuerdo en aceptar que, para todas las características, el excedente o defecto sobre el nivel exigido deba ser penalizado.

Con objeto de dar una solución a este planteamiento se propone la utilización de una generalización del "índice del máximo y mínimo nivel" creado por Jaume Gil Lafuente (2001), que además de poder penalizar solo los defectos, puede penalizar también solo los excesos.

El índice generalizado utilizará la distancia de Hamming y dos semidistancias para crear una ordenación entre el nivel de alejamiento de ciertos objetos y un objeto ideal.

A partir de la creación de tres grupos, tales que en el primero se agrupen las "k" características en que sí se penaliza el exceso o defecto sobre el nivel ideal, en el segundo, las "s" características en las que solo penaliza el exceso, y en el tercero, las "m" características en las que solo penaliza el defecto (k + s + m = n), se define el índice como:

$$I_g(x,Id) = \sum_{i=1}^{k} \left| Id_i - x_i \right| + \sum_{i=k+1}^{k+s} | \, mínimo(0; Id_i - x_i) \, | + \sum_{i=k+s+1}^{k+s+m} máximo(0; Id_i - x_i)$$

Evidentemente, en el caso de que todas las características no tengan el mismo grado de importancia, también se aconseja la ponderación mediante pesos de la siguiente manera:

$$I_g(x,Id) = \sum_{i=1}^{k} \left| Id_i - x_i \right| + \sum_{i=k+1}^{k+s} | \, mínimo(0; Id_i - x_i) \, | + \sum_{i=k+s+1}^{k+s+m} máximo(0; Id_i - x_i)$$

donde $\sum_{i=1}^{n} w_i = 1$

Veamos a continuación un ejemplo práctico de desarrollo de un *ranking* de posibles compañías aseguradoras a fin de entender las notaciones del índice.

Imaginemos cinco compañías A, B, C, D y E que ofrecen un seguro específico. Imaginemos seis características del mismo peso: C1, C2, C3, C4, C5 y C6, que un cliente considera esenciales para clasificar el interés que ofrece cada compañía. Supongamos que los grados de verdad que dicho cliente ha determinado para cada característica en cada compañía son:

	C1	C2	C3	C4	C5	C6
A	0,4	0,6	0,6	0,7	0,4	0,6
B	0,5	0,7	0,5	0,8	0,6	1
C	0,8	0,6	0,9	0,9	0,7	0,9
D	0,8	0,8	0,7	1	0,8	0,8
E	0,9	0,7	0,8	1	0,6	0,7

Supongamos también que el posible cliente ha determinado los valores ideales que considera óptimos de cumplir para cada característica.

	C1	C2	C3	C4	C5	C6
valor ideal	0.7	1	0.7	0.8	0.7	0.8

Finalmente, supongamos que el criterio de penalización para las dos primeras características es la distancia de Hamming, tanto si no se alcanza como si se sobrepasa el ideal; para las dos siguientes características se penaliza si no se consigue alcanzar el valor ideal, y en las dos últimas características se penaliza solo si se excede el valor ideal.

	|idi-xi|	|idi-xi|	|mín(0; idi-xi)|	|mín(0; idi-xi)|	máx(0; idi-xi)	máx(0; idi-xi)
A	|0,7-0,4| = 0,3	|0,7-0,6| = 0,4	|min(0;0,1)| = 0	|min(0;0,1| = 0	max(0;0,3) = 0,3	max(0;0,2) = 0,2
B	|0,7-0,5| = 0,2	|0,7-0,7| = 0,3	|min(0;0,2| = 0	|min(0;0| = 0	max(0;0,1) = 0,1	max(0;-0,2) = 0
C	|0,7-0,8| = 0,1	|0,7-0,6| = 0,4	|min(0;-0,2| = 0,2	|min(0;-0,1| = 0,1	max(0;0) = 0	max(0;-0,1) = 0
D	|0,7-0,8| = 0,1	|0,7-0,8| = 0,2	|min(0;0| = 0	|min(0;-0,2| = 0,2	max(0;-0,1) = 0	max(0;0) = 0
E	|0,7-0,9| = 0,2	|0,7-0,7| = 0,3	|min(0;-0,1| = 0,1	|min(0;-0,2| = 0,2	max(0;0,1) = 0,1	max(0;0,1) = 0,1

Consiguientemente, el índice generalizado de máximo y mínimo nivel para cada compañía es:

$$I_g\,(A,Id\,) = 0{,}3 + 0.4 + 0 + 0 + 0{,}3 + 0{,}2 = 1{,}2$$

$$I_g\,(B,Id\,) = 0{,}2 + 0.3 + 0 + 0 + 0{,}1 + 0 = 0{,}6$$

$$I_g\,(C,Id\,) = 0{,}1 + 0.4 + 0{,}2 + 0{,}1 + 0 + 0 = 0{,}8$$

$$I_g\,(D,Id\,) = 0{,}1 + 0.2 + 0 + 0{,}2 + 0 + 0 = 0{,}5$$

$$I_g\,(E,Id\,) = 0{,}2 + 0{,}3 + 0{,}1 + 0{,}2 + 0{,}1 + 0{,}1 = 1$$

A partir de los resultados podemos establecer la siguiente ordenación:

$$D \{ B \{ C \{ E \{ A$$

3. CONCLUSIÓN

El principal resultado que aporta el presente trabajo es la presentación de un índice que bajo supuestos de lógica borrosa permite obtener un modelo de ordenación de compañías aseguradoras.

Obviamente, para que el modelo sea útil en la praxis es necesario conocer las características que un consumidor espera que posea una compañía aseguradora. Por ello, se hace necesario identificar puntos fuertes y débiles empresariales utilizando la teoría de relaciones y grafos borrosos (Kaufmann, 1973), la teoría de efectos olvidados (Kaufmann y Gil Aluja, 1988) y la teoría de gestión de información de expertos (Kaufmann y Gil Aluja, 1993). Dicha metodología no se ha desarrollado en el presente trabajo por razones de espacio.

REFERENCIAS

Black, M. (1937). "Vagueness: An Exercise in Logical Analysis". *Philosophy of Science*, 4, 427-455.

Gil Lafuente, A.M. (1993). *El Análisis financiero en la incertidumbre*. Ariel.

Gil-Lafuente, J. (2001). *El "índice del máximo y mínimo nivel" en la optimización del fichaje de un deportista*. 10th AEDEM Int. Congress, 439-443.

Kaufmann, A. (1973). *Theory of Fuzzy Subsets*. Academic Press.

Kaufmann, A., y Gil Aluja, J. (1986). *Introducción de la teoría de los subconjuntos borrosos a la gestión de las empresas*. Editorial Milladoiro.

Kaufmann, A., y Gil Aluja, J. (1988). *Modelos para la investigación de efectos olvidados*. Editorial Milladoiro.

Kaufmann, A., y Gil Aluja, J. (1993). *Técnicas especiales para la gestión de expertos*. Editorial Milladoiro.

Linares, S., y Ferrer, J.C. (2014). "Teoría de conjuntos clásica vs teoría de subconjuntos borrosos. Un ejemplo elemental comparativo". *Tribuna Plural*, 1, 485-500.

Zadeh, L.A. (1965). "Fuzzy Sets". *Information and Control*, 8, 338-353.

58. STORM SURGE HAZARD ESTIMATION ON THE UNITED STATES' ATLANTIC COAST USING THE CLUSTERED SEPARATED. PEAKS-OVER-THRESHOLD SIMULATION (CSPS) METHOD

M. Lopeman
Columbia University, Department of Civil Engineering and Engineering Mechanics, USA,
mel2146@columbia.edu

G. Deodatis
Columbia University, Department of Civil Engineering and Engineering Mechanics, USA,
deodatis@civil.columbia.edu

G. Franco
Guy Carpenter & Company, Inc., UK,
Guillermo.E.Franco@guycarp.com

ABSTRACT

Coastal high water levels are becoming both higher and more frequent as global sea levels rise. Hazard estimation methods for storm surge return period computation vary widely across the literature and the choice of method greatly impacts results. Following the development of the clustered separated peaks-over-threshold simulation (CSPS) method for storm surge return period estimation in lower Manhattan, the authors of this method have extended the CSPS herein to each NOAA tidal gauge stations along the United States' eastern coast for which a sufficiently long time history of data is available. The CSPS is unique both in its separation of data into independent processes (tide, surge, and sea level rise) and its clustering method that accounts for surge duration. The result is a set of high water level return period estimates for locations distributed along the Atlantic coast, which can inform policy, insurance and engineering decisions.

1. INTRODUCTION

Damages from natural hazards such as earthquakes, tropical cyclones, droughts and wildfires have been increasing in the United States over the last 35 years (Smith and Katz, 2013). Furthermore, the distribution of said U.S. disasters over the period from 1980 to 2011 was dominated by tropical cyclone losses, rendering this peril of particular concern from a risk management perspective (Smith and Katz, 2013). Low-frequency, high-consequence events are particularly difficult to study because by definition, only a small amount of data occupies the extreme end of the tail. Despite this challenge, however, development and refinement of natural hazard estimation methodologies is an active and constantly evolving area of research because of its importance from a risk management perspective.

Methods for evaluating storm surge hazard —a potential source of damage associated with tropical cyclones— include physical approaches using finite element modeling of the effect of tropical cyclones on the ocean's coastal region as well as statistical approaches involving analysis of the frequency of high water levels observed in tidal gauge data (e.g. Lin *et al.*, 2012; Talke *et al.*, 2014; Pugh and Vassie, 1980). A recent instance of a tidal gauge data analysis procedure for computation of storm surge return periods is the clustered separated peaks-over-threshold simulation (CSPS) method (Lopeman *et al.*, 2014). Following the development of CSPS for estimation of water level return periods in lower Manhattan at the Battery, NY, the authors of the CSPS have extended the CSPS to all NOAA tidal gauge stations along the Atlantic coast and present the results herein.

2. CLUSTERED SEPARATED PEAKS-OVER-THRESHOLD SIMULATION METHOD

The CSPS computes return periods from t-years of hourly tidal gauge data through separation of the observations into three independent additive components-tidal cycle, sea level rise, and storm surge. Tidal cycle and sea level rise are treated deterministically and storm surge is treated stochastically using the peaks over threshold (POT) method. However, the CSPS uses a modified POT approach that clusters consecutive exceedances of the storm surge above some threshold height. Each exceedance cluster is approximated by a deterministic exponential rise-fall function,

whose parameters can be described stochastically. The resulting set of statistical parameters enables simulation of a longer artificial time series of water level data from which return periods are extracted. The length of the simulated time history N should be considerably longer than the highest desired return period to ensure and smooth results and convergence to the mean.

Benefits of the CSPS include its use of the generalized Pareto distribution, chosen for its ability to fit a wide tail with limited data, its inclusion of high water level data regardless of association with named tropical cyclones (in contrast to many physically-based methods), and its ability to compute return levels whose return periods are longer than the underlying tidal gauge time history (Lopeman *et al*., 2014). Lopeman *et al*. (2014) provide a case study at the Battery, NY illustrating that the CSPS produces lower return period estimates than others previously published for the same region; this suggesting that the CSPS may provide a more conservative representation of high water level exceedance probabilities than has previously been captured.

3. NOAA TIDAL GAUGE STATIONS

The National Oceanographic and Atmospheric Administration (NOAA) monitors hourly coastal water levels at more than 100 stations along the Eastern coast of the United States. The oldest tidal gauge station is station no. 8574680 located in Baltimore, MD and its digital hourly water records began in 1902. Many are much newer, having begun digital records as late as 2010. Some stations recorded monthly or even daily water levels much earlier than the hourly records began, but because the CSPS relies on hourly time histories, these earlier portions of the data are not useful in the CSPS. In this study, the CSPS was applied to all 30 NOAA tidal gauge stations along the Atlantic and Gulf Coasts whose digital hourly water level records are at least 30 years long. This distinction was made in order to ensure that estimates of the 50-year and 100-year water levels are reasonably reliable; tidal gauge time histories that are too short would not support extrapolation to higher return periods. The 30 NOAA stations to which the CSPS was applied are shown and listed in Figure 1.

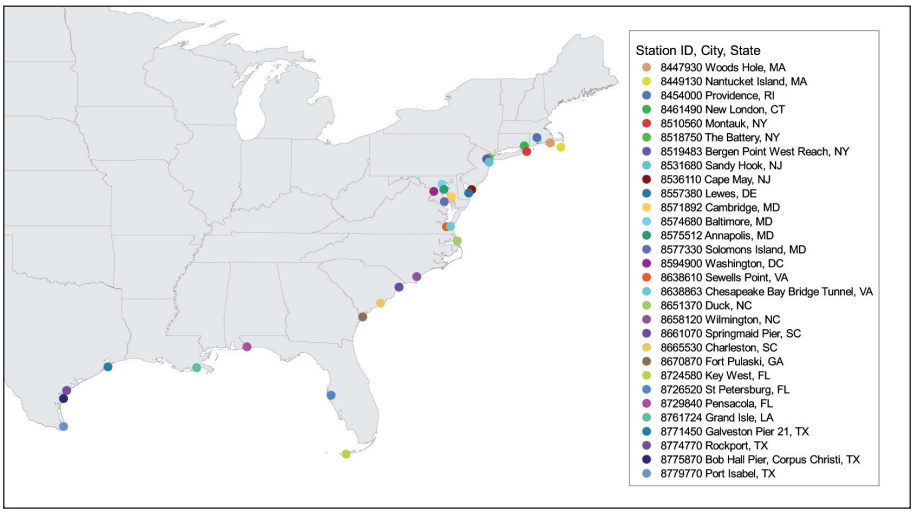


Figure 1. NOAA tidal gauge station locations along the Eastern coast of the United States. Only stations whose tidal gauge time histories are 30 years or longer are included.

To compute confidence intervals on the water levels, the bootstrapping algorithm developed with the CSPS was used at all 30 sites with 200 resamplings. This enabled the extraction of the 95% confidence interval at each desired return period. The length of the simulated time history for each station was chosen to be equal to 100 times the length of the original tidal gauge time history, or $N = 100t$. This was experimentally found to produce convergent and smooth water level curves at and below the 100-year return period at all tested stations.

4. RESULTS AND DISCUSSION

The results of the application of the CSPS to the tidal gauge stations consist of 30 sets of mean return levels and 95% confidence intervals for return periods less than or equal to 100t, the length of the simulated tidal gauge time history. For the sake of comparison, only the 10-year and 100-year water levels and confidence intervals are extracted. These return periods are selected because they characterize moderate and extreme hazard, respectively, and because NOAA computes water levels for the same return periods. NOAA's values are found by fitting the generalized extreme

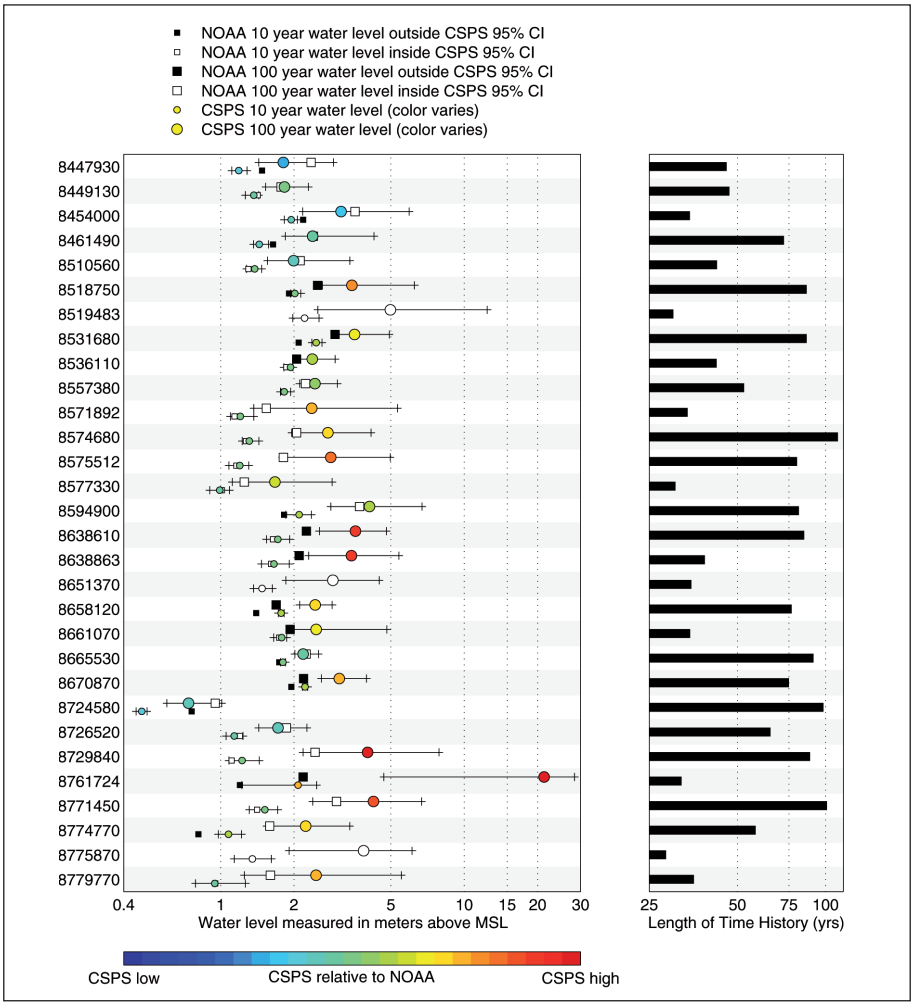


Figure 2. Comparison of 10-year and 100-year water levels provided by NOAA and com- puted using CSPS for 30 NOAA tidal gauge stations. The 10-year levels are plotted with smaller markers and the 100-year levels with larger markers. The circles represent the mean CSPS water levels and the squares represent the mean NOAA water levels. The 95% CI from the CSPS is included for both return periods and is represented by horizontal black lines. The NOAA return levels that fall within the 95% confidence bounds from the CSPS are plotted in white; those falling outside the 95% CI are plotted in black. The color of the circular mean CSPS water level coordinate corresponds to the discrepancy between the mean CSPS water level and the NOAA water level —when the CSPS mean is higher than NOAA's value, the marker is red, and when the CSPS mean is lower than NOAA's value, the marker is blue. For stations at which no NOAA return levels are available, the CSPS mean estimates are plotted in white. The right hand plot shows t, the length of the underlying tidal gauge time histories.

value distribution (GEV) to yearly maximum water levels at select tidal gauge stations (Zervas, 2013). The extreme water levels from NOAA are related, but not equal, to base flood elevations used by FEMA to map flood zones (Zervas, 2013).

Several key differences exist between the NOAA extreme levels and those computed from the CSPS. Firstly, NOAA does not separate tide and surge components of total water level, whereas the CSPS does. However, both methods do subtract sea level rise trend before analyzing extremes. Additionally, NOAA provides two sets of extreme water levels: one referenced to the mean sea level datum associated with the 1983-2001 tidal epoch, and another that extrapolates the mean sea level rise trend from this epoch forward in time to 2014. Because the CSPS results represent high water level hazard at the end of 2013, it is reasonable to compare the CSPS results to the 2014 NOAA water levels.

Comparison of NOAA's estimates to those obtained using the CSPS are provided in Figure 2. NOAA return levels are available for 27 out of the 30 stations studied[1]. The comparison suggests no particular trend in the relation of the CSPS and NOAA values; CSPS results are neither consistently lower nor consistently higher than those provided by NOAA. However, when CSPS results are higher than NOAA estimates, they tend to diverge much more than when CSPS results are lower than NOAA's, as evidenced by the preponderance of dark red circles compared to the lack of dark blue ones. Thus, the CSPS may produce overall higher return levels, especially when looking at higher return periods.

The highest 100-year return level estimate from the CSPS is for station 8761724, Grand Isle, LA. It appears to be somewhat of an outlier when compared to the rest of the CSPS results, but the fact that its tidal gauge time history is on the shorter end of the range at least partially accounts for this. For the most part, the 10-year and 100-year return levels fall within the 1-5 meter range, though some of the confidence intervals extend outside this window.

There is no strong geographical trend evident in the results, which suggests that the specifics of the datasets matter much more than location. As expected, stations with

[1] NOAA does not provide return level estimates for stations 8519483, 8651370, or 8775870.

longer time histories have narrower confidence intervals and stations with shorter time histories have wider confidence intervals. The NOAA values fall within the 95% confidence interval from the CSPS in the 15 out of 27 cases for the 10-year return period and in 18 out of 27 for the 100-year return period. There are two conflicting explanations for this. First, the fact that the confidence intervals for the 10-year return levels are narrower than those for the 100-year return levels decreases the likelihood that NOAA's estimate will fall within the bounds for the 10-year levels. Simultaneously, we expect less variability across methods for the 10-year return levels because in all cases, the underlying time histories are at least 30 years long. In other words, the higher the return period, the more likely divergence between methods. Thus, the numbers of NOAA values that fall within the 95% CI for both return periods are very close.

5. CONCLUSIONS

In this work, the CSPS was applied to all NOAA tidal gauge stations along the eastern coast of the United States for which sufficiently long time histories (at least 30 years) were available and water level return periods were calculated. Comparison to NOAA's exceedance probability levels suggests that the CSPS estimates higher return levels than NOAA. However, it is also shown that the NOAA values fall within the 95% CI from the CSPS for more than half of the stations tested. Divergence of the results for stations at which the NOAA value falls outside the 95% CI underscores the challenging nature of extreme hazard analysis. That the CSPS produces relatively high water levels aligns with the results of the paper presenting the CSPS that applies the method to lower Manhattan (Lopeman *et al.*, 2014). The results of this study suggest a potential opportunity for continued research into storm surge hazard and new thinking about how this estimation can impact policy, design and insurance decision-making.

REFERENCES

Lin, N., Emanuel, K., Oppenheimer, M., and Vanmarcke, E. (2012). "Physically Based Assessment of Hurricane Surge Threat under Climate Change". *Nature: Climate Change*, 2 (6), 462-467.

Lopeman, M., Deodatis, G., and Franco, G. (2014). *Extreme Storm Surge Hazard Estimation in Lower Manhattan – Clustered Separated Peaks-over-Threshold Simulation (CSPS) Method*. Manuscript submitted for review.

Pugh, D.T., and Vassie, J.M. (1980). "Applications of the Joint Probability Method for Extreme Sea Level Computations". *Proceedings of the Institution of Civil Engineers* (London). Part 1 – Design & Construction, 69 (2), 959-975.

Smith, A.B., and Katz, R.K. (2013). "US Billion-Dollar Weather and Climate Disasters: Data Sources, Trends, Accuracy and Biases". *Natural Hazards*, 67 (2), 387-410.

Talke, S.A., Orton, P., and Jay, D.A. (2014). "Increasing Storm Tides in New York Harbor, 1844-2013". *Geophysical Research Letters*, 41 (9), 3149-3155.

Zervas, C. (2013). *Extreme Water Levels of the United States 1893-2010*. NOAA Technical Report NOS CO-OPS 067.

59. MULTIVARIATE DENSITY AND DERIVATIVE GAMMA KERNEL ESTIMATION BY DEPENDENT DATA

L.A. Markovich
Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia,
Moscow Institute of Physics and Technology (MIPT) State University,
kimo1@mail.ru

ABSTRACT

We consider the nonparametric estimation of the multivariate probability density function and its partial derivative with a support on $[0, \infty)$ by dependent data. We use the class of gamma-kernel estimators which are asymmetric. The gamma kernels are nonnegative and change their shape depending on the position on the semi-axis. They possess good boundary properties for a wide class of densities useful in many applications like engineering, signal processing, actuarial science etc. The theoretical asymptotic properties of the multivariate density and its partial derivative estimates like biases, variances and covariances are derived. We obtain the optimal bandwidth selection for both estimates as a minimum of the mean integrated squared error (MISE) assuming dependent data with a strong mixing. Optimal rates of convergence of the MISE are found.

1. INTRODUCTION

Nonnegatively supported probability density functions (pdf) can be used to model a wide range of applications in engineering, signal processing, medical research, quality control, actuarial science and climatology among others. Regarding the optimal filtering in the signal processing and control of nonlinear processes the exponential pdf class is used, (Dobrovidov, 2012). Most total insurance claim distributions are shaped like gamma pdfs (Furman, 2008): nonnegatively supported , skewed

to the right and unimodal. The gamma distribution are also used to model rainfalls (Aksoy, 2000). Erlang and χ^2 pdfs are widely used in modeling insurance portfolios (Hürlimann, 2001).

The wide use of these pdfs in practice leads to the need of their estimation by finite data samples. One of the most common nonparametric pdf estimation methods are kernel estimators. The kernel density estimators were originally introduced for univariate independent identically distributed (iid) data using symmetrical kernels. The latter approach was widely adopted for tasks where a support of the underlying pdf f is unbounded. In case when the pdf has a nonsymmetric support $[0, \infty)$ the problem of the large bias on the zero boundary appears. This leads to a bad quality of the estimates. Obviously, the boundary bias of the multivariate pdf estimation becomes even more solid (Bouezmarni, Rombouts, 2007). To overcome this problem one can use special transformations as data reflection, the boundary kernels and others. Another solution is to use the asymmetrical kernel functions instead of symmetrical ones. For univariate nonnegative iid random variables (r.v.s), the estimators with gamma kernels were proposed in (Chen, 2000). The gamma kernel estimator was developed for univariate dependent data in (Bouezmarni, Rombouts, 2010). In (Bouezmarni, Rombouts, 2007) the gamma kernel estimator of the multivariate pdf was introduced for nonnegative iid r.v.s.

The gamma kernel is nonnegative and flexible regarding the shape. This allows to fit accuratly the multi-modal pdfs and their derivatives. Estimators constructed by gamma kernels have no boundary bias if $f''(0) = 0$ holds, i.e when the underlying density $f(x)$ has a shoulder at $x = 0$ (Zhang: 2010). This shoulder property is fulfilled, for example, for a wide exponential class of pdfs.

Along with the density estimation it is often necessary to estimate the derivative of the pdf. The estimation of the univariate density derivative by the gamma kernel estimator was proposed in (Dobrovidov, Markovich, 2013a, b) for iid data and in (Markovich, 2014) for a satisfying a strong mixing condition dependent data.

In this work we introduce the gamma product kernel estimators for a multivariate density with nonnegative support and its partial derivative by the multivariate dependent data with a strong mixing condition. The asymptotical behavior of the

estimates and the optimal bandwidths in the sense of minimal MISE are obtained. The derivative estimation requires a specific bandwidth different from that for the pdf estimation.

The pdf derivative could be useful to find the slope of the pdf curve, its local extremes, saddle points, significant features in data as well as in regression analysis (Brabanter, 2011). The pdf derivative also plays a key role in clustering via mode seeking (Sasaki, 2014). Furthermore, we are able to write down the estimator for the logarithmic derivative of the nonnegative pdf using multivariate dependent data. The estimate of the logarithmic derivative of the pdf is necessary for an optimal filtering in signal processing and control of nonlinear processes (Dobrovidov, 2012).

2. GAMMA KERNEL ESTIMATION

Let $\{\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T\}_{i=1}^n$ be a strongly stationary sequence of d-dimensional variables with an unknown probability density function $\{\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T\}_{i=1}^n$ which is defined on the bounded support $x_1^d \in \mathbf{R}^{+d}$. We assume that the sequence $\{\mathrm{X}_s\}$ is α–mixing with coefficient

$$\alpha(k) = \sup_t \sup_{\substack{A\in\sigma\{X_s, s\le t\} \\ B\in\sigma\{X_s, s>t+k\}}} |P(A\cap B) - P(A)P(B)|, \quad k \ge 1.$$

Here $\alpha(k) \to 0$ as $k \to \infty$. The α-mixing condition, also called strong mixing, is satisfied by many stochastic processes, i.e. autoregressive process (Donald, 1983). For these sequences we will use a notation $\{\mathrm{X}_s\} \in S(\alpha)$. To estimate the unknown pdf we use the product gamma kernel estimator

$$\widehat{f}(x_1^d) \;=\; \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K_{\rho(x_j,b_j),b_j}(X_{ij}), \tag{1}$$

where $b \equiv \{b_j\}_{j=1}^d$ is the vector of the bandwidth parameters such that $b \to 0$ as $n \to \infty$. As a kernel functions for each variable we propose the gamma kernel

$$K_{\rho(x,b),b}(t) = \frac{t^{\rho(x,b)-1}\exp(-t/b)}{b^{\rho(x,b)}\Gamma(\rho(x,b))}, \tag{2}$$

where $\Gamma(\cdot)$ is a standard gamma function. To improve the properties of the gamma kernel estimator Chen (2000) used a parameter *ρ(x, b)* defined as

$$\rho(x,b) \;=\; \begin{cases} \rho_1(x,b) = x/b, & \text{if} \quad x \geq 2b, \\ \rho_2(x,b) = (x/(2b))^2 + 1, & \text{if} \quad x \in [0, 2b). \end{cases} \tag{3}$$

Since a gamma kernel is nonnegative it has no weight on the negative semi- axes in contrast to symmetrical kernel estimators. Hence, the use of the gamma kernels is more natural for the estimation of the nonnegative supported pdfs. In (Dobrovidov, Markovich, 2013a; Markovich, 2014) the derivative of the univariate pdf was estimated just like the derivative of the gamma kernel estimator. For the multivariate case, we can analogically estimate any partial derivative of $f(x^n_{n-\tau})$. For example, the partial derivative of $f(x^n_{n-\tau})$ by x_n can be estimated as

$$\widehat{f}'_{x_n}(x^n_{n-\tau}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\prod_{j=n-\tau}^{n} K_{\rho(x_j,b_j),b_j}(X_{ij})\,(K_{\rho(x_n,b_n),b_n}(X_{in}))'_{x_n}, \tag{4}$$

where $(K_{\rho(x,b),b}(t))'_{xn}$ is the partial derivative of (2).

It is natural to consider the mean integrated squared errors as the measure of error of the proposed estimators. The unknown smoothing parameters are obtained as the minima of MISEs.

3. CONVERGENCE RATE OF THE DENSITY ESTIMATOR

We obtain the asymptotic properties of the pdf estimator. All proofs hold assuming that all the components of the bandwidth vector are different. For the practical use we give a simpler formulation of the lemmas for the equal bandwidths $b_1 = b_2 = \ldots = b_n = b$.

Lemma 3.1. If $b \to 0$ as $n \to \infty$, then the bias for the pdf estimate is equal to

$$Bias(\widehat{f}(x^n_{n-\tau})) \;=\; \frac{b}{2}\sum_{i=n-\tau}^{n} x_j \frac{\partial^2 f(x^n_{n-\tau})}{\partial x_j^2} + o(b). \tag{5}$$

Lemma 3.2. If $b \to 0$, $nb^{\frac{\tau+1}{2}} \to \infty$ as $n \to \infty$, then the variance expansion of the pdf estimate is equal to

$$\begin{aligned} var(\widehat{f}(x_{n-\tau}^n)) &= \frac{b^{-\frac{(\tau+1)}{2}}}{n}\left(\prod_{j=n-\tau}^{n}\frac{x_j^{-1/2}}{2\sqrt{\pi}}\right)\left(f(x_{n-\tau}^n)+bv_1(x_{n-\tau}^n)+b^2v_2(x_{n-\tau}^n)\right) \\ &- \frac{1}{n}\left(f(x_{n-\tau}^n)+\frac{1}{2}\sum_{j=n-\tau}^{n}x_jb\frac{\partial^2 f(x_{n-\tau}^n)}{\partial x_j^2}\right)^2+o(b^2), \end{aligned} \tag{6}$$

where

$$\begin{aligned} v_1(x_{n-\tau}^n) &= \sum_{j=n-\tau}^{n}\left(-\frac{1}{2}\frac{\partial f(x_{n-\tau}^n)}{\partial x_j}+\frac{x_j}{4}\frac{\partial^2 f(x_{n-\tau}^n)}{\partial x_j^2}\right), \\ v_2(x_{n-\tau}^n) &= -\sum_{j=n-\tau}^{n}\sum_{i=n-\tau}^{n}\frac{x_j}{8}\frac{\partial^3 f(x_{n-\tau}^n)}{\partial x_j^2\partial x_i}. \end{aligned} \tag{7}$$

Lemma 3.3. Let $\{X_j\}_{j\geq 1} \in S(\alpha)$ and $\int_1^\infty \alpha(\tau)^v d\tau < \infty, \quad 0 < v < 1$ hold, $f(x_{n-\tau}^n)$ be a twice continuously differentiable function, $b \to 0$ and $nb^{(\tau+1)\frac{v+1}{2}} - \infty$ as $n \to \infty$. If additionally there exists $M > 0$ such that for all $\tau > 1$ and $(x, y) \in \mathbb{R}^2$ it follows

$$|f_\tau(x,y) - f(x)f(y)| \leq M,$$

then the covariance of the multivariate pdf estimator is given by

$$\begin{aligned} |C(\widehat{f}(x_{n-\tau}^n)| &\leq \frac{2}{n}\sum_{\tau=2}^{c(n)} Cov\left(\widetilde{\widetilde{K}}\left(X_1,x,b\right),\widetilde{\widetilde{K}}\left(X_\tau,x,b\right)\right) \\ &+ \frac{2}{n}\sum_{\tau=c(n)+1}^{\infty} Cov\left(\widetilde{\widetilde{K}}\left(X_1,x,b\right),\widetilde{\widetilde{K}}\left(X_\tau,x,b\right)\right)| = I_1 + I_2 \end{aligned}$$

and its rate of convergence satisfies

$$\begin{aligned} I_1 &\leq \frac{2M}{nb^{\frac{\tau+1}{8}}}, \\ I_2 &\leq \frac{D(\kappa,x_{n-\tau}^n)}{nb^{\frac{(\tau+1)}{16}}}\left(f(x_{n-\tau}^n)\frac{6\kappa-1}{2(2\kappa-1)}+bS(\kappa,x_{n-\tau}^n)\right)^{1-2\kappa}\sum_{\tau=2}^{\infty}\tau\alpha(\tau)^{2\kappa}+o\left(b^2\right) \end{aligned} \tag{8}$$

Theorem 3.1. Assuming the conditions of Lemmas 3.1, 3.2, 3.3, the optimal bandwidth which provides the minimum of the MISE is given by

$$b = \left(\frac{(\tau+1)\int_0^\infty \left(\prod_{j=n-\tau}^{n} \frac{x_j^{-1/2}}{2\sqrt{\pi}} \right) f(x_{n-\tau}^n) dx_{n-\tau}^n}{\int_0^\infty \left(\sum_{j=n-\tau}^{n} x_j \frac{\partial^2 f(x_{n-\tau}^n)}{\partial x_j^2} \right)^2 dx_{n-\tau}^n} \right)^{\frac{2}{5+\tau}} n^{-\frac{2}{5+\tau}}. \tag{9}$$

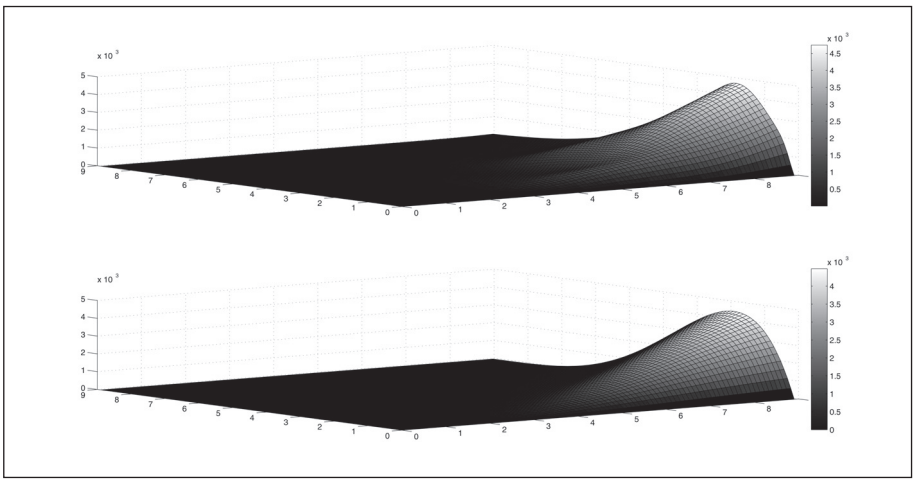


Figure 1. Gamma-kernel estimate of the pdf (top), true density (bottom).

The gamma-kernel estimate (1) of the McKay bivariate gamma pdf

$$f(x,y) = \frac{x^{\alpha-1}(y-x)^{\beta-1}\exp(-y/\mu)}{\mu^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)}, \quad x \le y, \quad y, \alpha, \beta > 0, \quad 0 < \mu < \infty$$

is shown on Figure 1 for $n = 2000$.

4. CONCLUSION

We define gamma kernel estimators of a multivariate density and its partial derivative. The bias, the variance and the covariance of the pdf estimate are obtained.

Using these results we derive the optimal bandwidth and the corresponding rate of the optimal MISE. The same is done for the estimate of the partial pdf derivative in Markovich (2015).

REFERENCES

Aksoy, H. (2000). "Use of Gamma Distribution in Hydrological Analysis". *Turk J. Engin Environ Sci*, 24, 419 - 428.

Bouezmarni, T., Rombouts, J.V.K. (2007). "Nonparametric Density Estimation for Multivariate Bounded Data". *J. Statist. Plann. Inference,* 140(1), 139-152.

Bouezmarni, T., and Rombouts, J.V.K. (2010). "Nonparametric Density Estimation for Positive Times Series". *Comput. Statist. Data Anal.*, 54(2), 245-261.

De Brabanter, K., De Brabanter, J., and De Moor, B. (2011). "Nonparametric Derivative Estimation". *Proc. of the 23rd Benelux Conference on Artificial Intelligence (BNAIC)*, Gent, Belgium, 75-81.

Dobrovidov, A.V., Koshkin, G.M., and Vasiliev, V. A. (2012). *Non-Parametric State Space Models*. USA: Kendrick press.

Furman, E. (2008). "On a Multivariate Gamma Distribution". *Statist. Probab. Lett.*, 78, 2353-2360.

Hürlimann, W. (2001). "Analytical Evaluation of Economic Risk Capital for Portfolios of Gamma Risks". *ASTIN Bulletin*, 31, 107-122.

Markovich, L.A., and Dobrovidov, A.V. (2013). "Nonparametric Gamma Kernel Estimators of Density Derivatives on Positive Semi-Axis". *Proc. of IFAC MIM 2013*, 944-949, Petersburg, Russia.

Markovich, L.A., and Dobrovidov, A.V. (2013). "Data-Driven Bandwidth Choice for Gamma Kernel Estimates of Density Derivatives on the Positive Semi-Axis". *Proc. of*

IFAC International Workshop on Adaptation and Learning in Control and Signal Processing, 500-505, Caen, France

Markovich, L.A. (2014). "Gamma Kernel Estimation of the Density Derivative on the Positive Semi-Axis by Dependent Data". Submitted to *Revstat*.

Markovich, L.A. (2015). *Gamma Kernel Estimation of Multivariate Density and its Derivative on the Nonnegative Semi-Axis by Dependent Data* (working paper) (arXiv:1410.2507).

Sasaki, H.A., Hyvärinen, A., and Sugiyama, M. (2014). "Clustering via Mode Seeking by Direct Estimation of the Gradient of a Log-Density". *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML/ PKDD 2014), to appear.

Song Xi, Ch. (2000). "Probability Density Function Estimation Using Gamma Kernels". *Ann. Inst. Statist. Math*, 54, 471-480.

Zhang, S. (2010). "A Note on the Performance of the Gamma Kernel Estimators at the Boundary". *Statis. Probab. Lett.*, 80, 548-557.

60. HITTING TIMES OF THRESHOLD EXCEEDANCES AND THEIR DISTRIBUTIONS

N. Markovich
Institute of Control Sciences, Russian Academy of Sciences, Russia,
nat.markovich@gmail.com
markovic@ipu.rssi.ru

ABSTRACT

We investigate exceedances of the process over a sufficiently high threshold. The exceedances determine the risk of hazardous events like climate catastrophes, huge insurance claims, the loss and delay in telecommunication networks. Due to dependence such exceedances tend to occur in clusters. Cluster structure of social networks is caused by dependence (social relationships and interests) between nodes and possibly heavy-tailed distributions of the node degrees. A minimal time to reach a large node determines the first hitting time. We derive asymptotically equivalent distribution and a limit expectation of the first hitting time to exceed the threshold u_n as sample size n tends to infinity. The results can be extended to the second and, generally, to kth ($k > 2$) hitting times.

1. INTRODUCTION

Let $\{X_n\}_{n\geq 1}$ be a stationary process with marginal distribution function $F(x)$ and $Mn = max\{X_1, \ldots, X_n\}$. We investigate the exceedances of the process over a sufficiently high threshold $_n$. Due to dependence such exceedances tend to occur in clusters. Let us consider the inter-cluster size

$$T_1(u) = \min\{j \geq 1 : M_{1,j} \leq u, X_{j+1} > u | X_1 > u\}, \quad (1)$$

i.e. the number of inter-arrivals of observations running under the threshold between two consecutive exceedances, where $M_{1,j} = \max\{X_2, \ldots, X_j\}$, $M_{1,1} = -\infty$. Let $T^*(u_n)$ be a first hitting time to exceed the threshold u_n. We get

$$P\{T^*(u_n) = j+1\} \quad = \quad P\{M_j \le u_n, X_{j+1} > u_n\}, \tag{2}$$

$j = 0, 1, 2, \ldots, M_0 = -\infty.$

The necessity to evaluate quantiles of the first hitting time and its mean is arisen in many applications. In social networks it is important to compare sampling strategies (Avrachenkov *et al.*, 2012; Avrachenkov *et al.*, 2014a, 2014b; Chul-Ho Lee *et al.*, 2012) like random walks, Metropolis-Hastings Markov chains, Page Ranks and other with regard to how quickly they allow to reach a node with a large degree that is the number of links with other nodes. It is important to investigate the first hitting time of significant nodes since it allows us to disseminate advertisement or collect opinions more effectively within the clusters surrounded such nodes. Similar problem is required in telecommunication peer-to-peer networks to find a node with a large number of peers, Dán and Fodor (2009), Markovich (2013).

Definition 1.1. *Leadbetter et al., (1983, p. 53).* The stationary sequence $\{X_n\}_{n\ge1}$ is said to have extremal index $\theta \in [0, 1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau \qquad \textit{and} \tag{3}$$

$$\lim_{n\to\infty} P\{M_n \le u_n\} = e^{-\tau\theta} \tag{4}$$

hold.

The extremal index θ of $\{X_n\}$ relates to the first hitting time $T^*(u_n)$, Roberts *et al.* (2006). Really, since un is selected according to (3) it follows that $P\{X_n > u_n\}$ is asymptotically equivalent to $1/n$. Notice, that $P\{M_k \le u_n\} = P\{T^*(u_n) > k\}$. Hence, substituting τ by (3) we get from (4)

$$P\{T^*(u_n)/n > k/n\} \sim e^{-\theta k P\{X_n > u_n\}} \sim e^{-\theta k/n},$$

$$\lim_{n\to\infty} P(T^*(u_n)/n > x) = e^{-\theta x}$$

for positive x. It follows

$$\lim_{n\to\infty} E(T^*(u_n)/n) = 1/\theta. \tag{5}$$

This implies, the smaller θ, the longer it takes to reach an observation with a large value. Using achievements regarding the asymptotically equivalent geometric distribution of $T_1(x_{\rho n})$ derived in (Theorem 2, Markovich, 2014), where the $(1-\rho_n)$ th quantile $x_{\rho n}$ of $\{X_n\}$ is taken as u_n, we derive an asymptotically equivalent distribution of the first hitting time and its limiting expectation that specifies (5) in Section 2.

We use the mixing condition proposed in Ferro and Segers (2003)

$$\alpha_{n,q}(u) = \max_{1\le k\le n-q} \sup |P(B|A) - P(B)| = o(1), \qquad n\to\infty, \tag{6}$$

where for real u and integers $1 \le k \le l$, $F_{k,l}(u)$ is the σ-field generated by the events $\{X_i > u\}$, $k \le i \le l$ and the supremum is taken over all $A \in F_{1,k}(u)$ with $P(A) > 0$ and $B \in F_{k+q,n}(u)$ and k, q are positive integers.

The mixing condition used in Theorem 2, Markovich (2014), is simplified in Markovich (2015).

Theorem 1.1. Markovich (2015). Let $\{X_n\}_{n\ge1}$ be a stationary process with the extremal index θ. Let $\{x_{\rho n}\}$ and $\{x_{\rho^* n}\}$ be sequences of quantiles of X_1 of the levels $\{1-\rho_n\}$ and $\{1-\rho_{*n}\}$, respectively[1], those satisfy the conditions (3) and (4) if u_n is replaced by $x_{\rho n}$ or by $x_{\rho^* n}$ and, $q_n = 1-\rho_n$, $q_n^* = 1-\rho_n^*$, $\rho_n^* = (1-q_n^\theta)^{1/\theta}$. Let $\{q_{n,1}^*\}$ be a sequence of positive integers such that $q_{n,1}^* = o(j)$ for $j \ge j_0(n)$, $j_0(n) \to \infty$ as $n \to \infty$ and

$$\alpha_{j+1,q_{n,1}^*} = o(\rho_n) \tag{7}$$

holds as $n \to \infty$, where $\alpha_{n,q} = \alpha_{n,q}(x_{\rho n})$ is determined by (6). Then it follows for $j \ge j_0(n)$

$$\lim_{n\to\infty} P\{T_1(x_{\rho_n}) = j\}/(\rho_n(1-\rho_n)^{(j-1)\theta}) = 1. \tag{8}$$

[1] $\overline{F}(x_{\rho_n}) = P\{X_1 > x_{\rho_n}\} = \rho_n$.

The achievements can be similarly extended to the second and, generally, to kth ($k > 2$) hitting times. The paper is organized as follows. In Section 2 we derive the asymptotically equivalent distribution and the limit expectation of the first hitting time to exceed sufficiently high threshold. Proofs are given in Section 3.

2. DISTRIBUTION AND EXPECTATION OF THE FIRST HITTING TIME

Theorem 2.1. Let all conditions of Theorem 1.1 be satisfied. We assume

$$\sup_n \rho_n E(T_1(x_{\rho_n})) < \infty. \tag{9}$$

$$\lim_{n\to\infty} nP\{T^*(u_n) = n\} \;=\; e^{-\theta\tau}/\theta, \tag{10}$$

$$P\{T^*(x_{\rho_n}) = j\} \;\sim\; \frac{\rho_n}{\theta}(1-\theta\rho_n)^{j-1}. \tag{11}$$

for $j \geq j_0(n), j_0(n) \to \infty$ as $n \to \infty$.

Expression (11) implies that $P\{T^*(x_{\rho n}) = j\} \sim P\{T_\theta = \mathrm{j}\}$, where

$$T_\theta \;=\; \begin{cases} \chi, & \text{with probability} \quad 1/\theta^2; \\ 0, & \text{with probability} \quad 1-1/\theta^2 \end{cases}$$

holds and χ has the geometric distribution with probability $\rho_n\theta$.

Remark 2.1. The condition (9) provides a uniform convergence of the range $\sum_{j=1}^{\infty} P\{T_1(x_{\rho_n}) = j\}$ by n. The latter condition is fulfilled for a geometrically distributed $T_1(x_{\rho n})$, i.e. $P\{T_1(x_{\rho n}) = j\} = \rho_n(1-\rho_n)^{j-1}$.

Lemma 2.1. Let the conditions of Theorem 2.1 be satisfied. Then it follows

$$\lim_{n\to\infty} \rho_n E T^*(x_{\rho_n}) = 1/\theta^3, \tag{12}$$

where $E(T^*(x_{\rho n}))$ denotes $\sum_{j=j_0(n)}^{\infty} jP\{T^*(x_{\rho_n}) = j\}$.

Since $\rho_n \sim \tau/n$ according to (3), the expression (12) specifies (5).

3. PROOFS

3.1. Proof of Theorem 2.1

It follows from (2) that

$$\begin{aligned} P\{T^*(u_n) = n\} &= P\{M_{n-1} \le u_n, X_n > u_n\} \\ &= P\{M_{n-1} \le u_n\} - P\{M_n \le u_n\}. \end{aligned} \tag{13}$$

We obtain due to (1) and the stationarity of {Xn} that it holds

$$\begin{aligned} & P\{T_1(u_n) = n\} = P\{M_{1,n} \le u_n, X_{n+1} > u_n | X_1 > u_n\} \\ = \; & \left(P\{M_{1,n} \le u_n, X_{n+1} > u_n\} - P\{M_n \le u_n, X_{n+1} > u_n\}\right)/P\{X_1 > u_n\} \\ = \; & \left(P\{M_{n-1} \le u_n, X_{n+1} > u_n\} - P\{M_n \le u_n, X_{n+1} > u_n\}\right)/P\{X_1 > u_n\} \\ = \; & \left(P\{T^*(u_n) = n\} - P\{T^*(u_n) = n+1\}\right)/P\{X_1 > u_n\}. \end{aligned} \tag{14}$$

Following Ferro and Segers (2003) we get alternatively for $n \ge 1$

$$\begin{aligned} P\{T_1(u_n) > n\} &= P\{M_{1,n+1} \le u_n | X_1 > u_n\} \\ &= \left(P\{M_{1,n+1} \le u_n\} - P\{M_{n+1} \le u_n\}\right)/P\{X_1 > u_n\} \\ &= \left(P\{M_n \le u_n\} - P\{M_{n+1} \le u_n\}\right)/P\{X_1 > u_n\} \\ &= P\{T^*(u_n) = n+1\}/P\{X_1 > u_n\}. \end{aligned}$$

Thus, we get

$$\begin{aligned} P\{T^*(x_{\rho_n}) = n+1\} &= P\{X_1 > x_{\rho_n}\} \cdot P\{T_1(x_{\rho_n}) > n\} \\ &= \rho_n \sum_{i=n+1}^{\infty} P\{T_1(x_{\rho_n}) = i\}, \end{aligned} \tag{15}$$

where $\rho_n = P\{X_1 > x_{\rho n}\}$. Now we use the fact followed from (8) in Theorem 1.1, i.e. it holds

$$c_n P\{T_1(x_{\rho_n}) = i\} \sim \eta_n(1-\eta_n)^{i-1} \tag{16}$$

as $n \to \infty$, where $1-\eta_n = (1-\rho_n)^\theta$, $c_n = \eta_n/(1-(1-\eta_n)^{1/\theta})$, $0 < \eta_n < 1$. Furthermore, we obtain

$$\begin{aligned} P\{T^*(x_{\rho_n}) = n+1\} &= \frac{\rho_n}{c_n}\sum_{i=n+1}^{\infty} c_n P\{T_1(u_n) = i\} \sim \frac{\rho_n}{c_n}\sum_{i=n+1}^{\infty}\eta_n(1-\eta_n)^{i-1} \\ &= \frac{\rho_n}{c_n}(1-\eta_n)^n = \frac{\rho_n}{\eta_n}\left(1-(1-\eta_n)^{1/\theta}\right)(1-\eta_n)^n \\ &\sim \frac{\rho_n^2(1-\theta\rho_n)^n}{1-(1-\rho_n\theta)} = \frac{\rho_n}{\theta}(1-\theta\rho_n)^n \sim \frac{\rho_n}{\theta}e^{-\theta\tau}, \end{aligned} \tag{17}$$

since $(1-\rho_n)^\theta = 1-\theta\rho_n + o(\rho_n)$ and $\lim_{n\to\infty}(1-\theta\rho_n)^n = e^{-\theta\tau}$ as $\rho_n \sim \tau/n$, $n \to \infty$. Formula (17 implies both (10) rewriting it with regard to $P\{T^*(x_{\rho n}) = n+1\}/\rho n$ and (11) replacing $n+1$ by j.

Hence, we have from (17)

$$P\{T^*(x_{\rho_n}) = j\} \sim \frac{1}{\theta^2}\rho_n\theta(1-\rho_n\theta)^{j-1}, \qquad n \to \infty. \tag{18}$$

The similarity in the first string of (17) follows from (9) and since

$$\begin{aligned} \sup_n r_k(n) &= \sup_n \sum_{i=\lfloor k/\rho_n \rfloor}^{\infty} P\{T_1(x_{\rho_n}) = i\} \\ &= \sup_n \sum_{i=\lfloor k/\rho_n \rfloor}^{\infty} \frac{i}{i} P\{T_1(x_{\rho_n}) = i\} \le \sup_n \frac{\rho_n}{k} E(T_1(x_{\rho_n})) \to 0 \end{aligned}$$

holds as $k \to \infty$.

3.2. Proof of Lemma 2.1

Using (15) and (16) we find for $j_0(n) = o(k)$

$$\begin{aligned}
&\sum_{j=j_0(n)}^{k} jP\{T^*(x_{\rho_n}) = j\} = \sum_{j=j_0(n)}^{k} j\rho_n \sum_{i=j}^{\infty} P\{T_1(x_{\rho_n}) = i\} \\
\sim\ &\sum_{j=j_0(n)}^{k} j\frac{\rho_n}{c_n}\sum_{i=j}^{\infty}\eta_n(1-\eta_n)^{i-1} = \frac{\rho_n}{c_n\eta_n}\sum_{j=j_0(n)}^{k} j\eta_n(1-\eta_n)^{j-1} \\
=\ &\frac{\rho_n}{c_n\eta_n}\left(\sum_{j=1}^{k} j\eta_n(1-\eta_n)^{j-1} - \sum_{j=1}^{j_0(n)-1} j\eta_n(1-\eta_n)^{j-1}\right) \\
\sim\ &\frac{\rho_n}{c_n\eta_n}\sum_{j=1}^{k} j\eta_n(1-\eta_n)^{j-1} \to \frac{\rho_n}{c_n\eta_n}\sum_{j=1}^{\infty} j\eta_n(1-\eta_n)^{j-1}
\end{aligned}$$

as k $\to \infty$. We obtain the expectation of the first hitting time

$$\begin{aligned}
ET^*(x_{\rho_n}) \ &\sim\ \frac{\rho_n}{c_n\eta_n}\sum_{j=1}^{\infty} j\eta_n(1-\eta_n)^{j-1} \\
&=\ \frac{\rho_n^2}{\eta_n^3} = \frac{\rho_n^2}{(1-(1-\rho_n)^\theta)^3} \sim \frac{\rho_n^2}{\rho_n^3\theta^3} = \frac{1}{\rho_n\theta^3}.
\end{aligned}$$

The similarity follows from the same arguments as before. Then (12) follows.

ACKNOWLEDGMENTS

Special thanks to the partial financial support by the Russian Foundation for Basic Research, grant 13-08-00744 A.

REFERENCES

Avrachenkov, K., Litvak, N., Sokol, M., and Towsley, D. (2012). "Algorithms and Models for the Web Graph". Lecture notes in *Computer Science*, 7323, *Quick Detection of Nodes with Large Degrees*, Springer Berlin Heidelberg, 54-65.

Avrachenkov, K., Markovich, N., and Sreedharan, J.K. (2014a). "Distribution and Dependence of Extremes in Network Sampling Processes". *Research report no. 8578, INRIA*, http://hal.inria.fr/hal-01054929

Avrachenkov, K., Markovich, N., and Sreedharan, J.K. (2014b). "Distribution and Dependence of Extremes in Network Sampling Processes. Workshop on Complex Networks and their Applications". *The 10th IEEE International Conference on Signal-Image Technology and Internet-Based Systems* (SITIS 2014) November 23-27, 2014, Marrakech, Morocco, 331-338.

Chul-Ho Lee, Xin Xu, and Do Young Eun (2012). "Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling". *CoRR*.

Dán, G., and Fodor, V. (2009). "Delay Asymptotics and Scalability for Peer-to-Peer Live Streaming". *IEEE Trans. Parallel Distrib.*, 20 (10), 1499Ű1511.

Ferro, C.A.T., and Segers, J. (2003). "Inference for Clusters of Extreme Values". *J. R. Statist. Soc.* B., 65, 545-556.

Leadbetter, M.R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequence and Processes*. Springer, New York.

Markovich, N.M. (2013). "Quality Assessment of the Packet Transport of Peer-to-Peer Video Traffic in High-Speed Networks". *Performance Evaluation*, 70, 28U˝ 44.

Markovich, N.M. (2014). "Modeling Clusters of Extreme Values". *Extremes*, 17 (1), 97-125.

Markovich, N.M. (2015). *Clusters of Extremes: Models, Estimation and Applications* (submitted).

Roberts, G.O., Rosenthal, J. S., Segers, J., and Sousa, B. (2006). "Extremal Indices, Geometric Ergodicity of Markov Chains and MCMC". *Scandinavian Journal of Statistics*, 9, 213-229.

61. INTERPRETING TEXTUAL USER GENERATED CONTENT FOR SUPPLIER-CREDIT RISK ANALYSIS

A. Mengelkamp, S. Hobert and M. Schumann
University of Goettingen, Germany,
amengel@uni-goettingen.de
shobert@uni-goettingen.de
mschuma1@uni-goettingen.de

ABSTRACT

Irrecoverable receivables due to insolvencies of debtors can jeopardize the own liquidity. Therefore supplier-credit risk analysis should be continuously improved. We investigate how user generated content (UGC) from social media platforms needs to be interpreted in order to improve the information base for supplier-credit risk analysis. We compare data from the microblogging platform Twitter related to solvent and insolvent companies. Our results indicate that the amount of UGC for insolvent companies is greater prior to their insolvency. Furthermore, it can be distinguished between solvent and insolvent companies based on the sentiment in UGC. Managers promote companies before an insolvency. They engage in restructuring and expansion endeavours. Beneficial cooperations and potential investors are also discussed widely in UGC before an insolvency.

1. INTRODUCTION

In order to prevent irrecoverable receivables arising due to insolvent debtors which can jeopardize the own liquidity, lenders in a credit relationship attempt to give credit to borrowers who are likely to pay their debts only (Schumann, 2002). Nevertheless, the lender has imperfect information concerning the actions the borrower has undertaken or will undertake in the future. This asymmetry of information is

especially great for supplier-credits because lenders cannot access information from regular accounts in contrast to banks (Beck, 2014). The asymmetry of information is reduced during the process of credit assessment (Stiglitz, 1988). Hereby, the likelihood of payment default is predicted based on statistical analysis of former credit agreements and a broad basis of information (Helfrich, 2010). This can encompass annual reports, previous payment behaviour as well as industry and management analyses in supplier-credit relationships (Schumann, 2002). The procedures of credit assessment should be continuously analysed and improved in order to diminish bad debt (Kramer & Nitsch, 2010). Bartels (2013) suggests that data from social media platforms which build on the ideological and technological foundations of the Web 2.0 and allow the exchange, reuse and remix of UGC should be integrated in the information base for credit risk analysis and unveils that a research gap is present in this domain (Kaplan & Haenlein, 2010). Furthermore, it has been shown, that early indicators from UGC, which "can be seen as the sum of all ways in which people make use of social media" (Kaplan & Haenlein, 2010, p. 61), can predict changes in economic and commercial indicators (Bollen, Mao and Zeng, 2011). Therefore we want to assess:

How can UGC be interpreted in order to reduce the information asymmetry between the lender and the borrower in a supplier-credit relationship?

We start with a presentation of our set of textual data. Then, we describe how we analysed the data, present our findings and discuss them. At the end we unfold limitations and complete our paper with a conclusion.

2. DATA COLLECTION

We refer to a statistic published by the German Federal Bureau of Statistics which discloses the ten biggest German insolvencies in the year 2013 measured by the number of affected employees (Federal Bureau of Statistics, 2014). We expect a great amount of UGC due to the high number of people affected. Then, we determined a solvent company for each insolvent company to build a reference group. Hereby, we accessed the Hoppenstedt company database provided by Bisnode, which is a European vendor for economic information. For each of the insolvent

companies we selected the company with the most compliant number of employees which has the same legal form (stock company/company with limited liability) and operates in the same industry according to the latest classification of industries in Germany (Federal Bureau of Statistics, 2008). In addition to that, we discussed industry analyses in order to verify the affiliation of companies to industries beyond the official classification.

As a source of UGC we selected the major microblogging platform Twitter since it has been proven that valuable indicators for changes in economic conditions can be extracted from the posts which are called Tweets (Bollen *et al.,* 2011). We implemented a crawler to extract all German Tweets for each pair of companies in the year prior to the date on which the insolvent company filed for bankruptcy. As search terms we curtailed the official names of the companies insofar as they still allow a distinct matching of Tweets to the respected businesses. This was done in order to capture relevant Tweets even when the full name was not mentioned by the authors. An overview of the resulting dataset which consists of 7071 Tweets can be seen in Table 1.

Insolvent companies		Solvent companies	
Company name	Number of Tweet	Number of Tweet	Company Name
Praktiker Deutschland GmbH	198	216	Obi GmbH
Walter Services GmbH	367	169	Arvato direct services GmbH
Baumarkt Max Bahr GmbH & Co. KG	1201	139	Hagebaumarkt GmbH & Co. AG
Kunert Fashion GmbH & Co. KG	6	19	Bugatti GmbH
Conergy AG	214	264	Centrotherm Photovoltaics AG
Alpine Bau GmbH	393	101	Hochtief Solutions AG
Loewe AG	383	105	Medion AG
Backstube Siebrecht GmbH & Co. OHG	30	4	Bäcker Görtz GmbH
Flexstrom AG	2439	162	Badenova AG & Co. KG
IVG Immobilien AG	555	106	PlanetHome AG

Table 1. Overview of dataset.

3. DATA ANALYSIS AND DISCUSSION

For data analysis we followed the methodological approach of content analysis described by Neuendorf (2002). We decided to code our data manually since previous research does not focus on the domain of credit risk analysis and well established sentiment dictionaries for our purpose in the German language do not exist. We labeled each tweet as described in Table 2 based on suggestions by Neuendorf (2002).

Score	Requirement
-2	Tweet contains evidence concerning financial instability regarding the company
-1	Tweet does not contain evidence of financial instability but sentiment is negative regarding the company
0	Tweet does not contain sentiment regarding the company
1	Tweet does not contain evidence of financial stability but sentiment is positive regarding the company
2	Tweet contains evidence concerning financial stability regarding the company

Table 2. Codebook.

The coding was done by two researchers independently who achieved an intercoder reliability of 93.81%. During a joint verification the coders agreed on a final coding. The majority of all tweets (5786, 81,8%) is related to insolvent companies. Only 1285 (18,2%) are related to solvent companies as seen in Table 3. A negative weight was assigned to 53,7% of Tweets related to insolvent companies whereas 36,6% are neutral and only 9,7% contain positive sentiment. Concerning solvent companies 79,5% are neutral and a minority of 5,8% is negative, respectively 14,7% are positive. This indicates that negative sentiment can be associated with financial instability. Nevertheless, positive sentiment is not necessarily related to financial stability. Since almost 80% of Tweets related to solvent companies contain no sentiment, neutral UGC is most likely associated with financial stability. In the following we analyse the 563 positive Tweets related to insolvent companies and the 74 negative Tweets related to solvent companies.

	Codes					
	-2	-1	0	1	2	Sum
Insolvent companies	2747 (47,5 %)	359 (6,2 %)	2117 (36,6 %)	458 (7,9 %)	105 (1,8 %)	5786 (81,8 %)
Solvent companies	40 (3,1 %)	34 (2,7 %)	1022 (79,5 %)	100 (7,8 %)	89 (6,9 %)	1285 (18,2 %)

Table 3. Distribution of Tweets.

We do this in order to investigate what positive aspects are discussed prior to an insolvency and what negative aspects are discussed for solvent companies. We rely on the methodological approach for grounded theory of Glaser (1978) for deeper analysis of the mentioned Tweets. Hereby we assigned categories, which describe the covered topics, to each Tweet. We identified the categories listed in Table 4 and Table 5 through an iterative process of classification and subsequent discussion.

It is evident that positive news related to a company prior to its bankruptcy are published even when they occur in a negative context. We labelled these statements as issued by company internals when this was obviously done. For example, executives from Praktiker, a major German home improvement market, claim, that they were able to diminish the decline of turnover prior to the insolvency. Furthermore, executives from Flexstrom declared that they are very satisfied with the past business year and that Flexstrom does not have financial problems in the year prior to the insolvency. We labelled statements as positive news or positive news in a negative context, when the origin could not be identified. An example for positive news in a negative context are Tweets in which is stated that Loewe, which is a producer of electronics, has a schedule for getting out of the red figures. Discussions about restructuring processes, the expansion of business activities, cooperations with other companies or capital enhancements can also indicate that a company is not performing well since these topics occurred in Tweets prior to insolvencies.

Topic	Occurences
Positive news in negative context	144
Positive news	104
Investor buys into company	76
Expansion of company	59
Cooperation with other companies	53
Positive statement from company-internals	44
Restructuring	41
Capital increase	19
Order obtained	9
Positive Speculation	8
Potential customers	6

Table 4. Positive Tweets related to insolvent companies.

For further analysis of negative Tweets related to solvent companies as seen in Table 5 it is important to know that all but three Tweets are related to Centrotherm Photovoltaics AG. This company is operating in the German solar industry which is subject to a major crisis (Fücks, 2013). Centrotherm Photovoltaics was undergoing a protective-screen procedure in which German companies can recapitalize themselves and prevent a filing under regular insolvency law. Centrotherm Photovoltaics was restructured successfully. One of the remaining three negative tweets is related to Hagebaumarkt, a German home improvement store. In the Tweet the shutdown of a store is mentioned. The other two Tweets report that the German electric company Badenova AG could not benefit from the ongoing coldness and was subject to a power failure. This indicates that almost no negative Tweets are related to well performing companies. However, prior to insolvencies the great number of negative information is enriched with discussions about undergoing positive changes in the companies and positive highlighting partially from company-internals.

Topic	Occurences
Restructuring	31
Protective screen procedure	24
Order resigned	14
Negative news	2
Unsatisfied employees	1
Restriction of business activities	1
Unsatisfied customers	1

Table 5. Negative Tweets related to solvent companies.

5. LIMITATIONS

Our research is limited by size and scope of our sample. It has to be investigated if similar or additional results can be obtained with different companies and other sources of UGC. Furthermore, probably not all relevant Tweets were extracted due to the search terms. We also restricted our study to German Tweets within one year prior to insolvency. UGC in other languages and a longer timespan might generate different results. Although we verified the coding of the Tweets by two researchers, other coders might come up with different results.

6. CONCLUSION

We are able to show that valuable information can be extracted from UGC before an insolvency. It is evident that executives try to promote their company before an insolvency in contrast to the negative UGC. A restructuring of processes is emphasized and beneficial cooperations with other companies as well as investors are highlighted.

This knowledge could be used in supplier as well as other forms of credit risk analysis in order to reduce the asymmetry of information between the lender and the borrower. Our research approach needs to be continued using quantitative statistics in order to be able to build models for decision support of credit risk analysis and evaluate if negative company performance can be predicted earlier than before.

REFERENCES

Bartels, J.C. (2013). "Credit Scoring in the Cross-hairs of Consumerism". *Password,* 28 (4), 9.

Beck, T.R. (2014). *Inkassounternehmen und der Erfolg beim Forderungseinzug.* Wiesbaden: Springer Gabler.

Bollen, J., Mao, H., and Zeng, X.-J. (2011). "Twitter Mood Predicts the Stock Market". *Journal of Computational Science,* 2 (1), 1-8.

Federal Bureau of Statistics (2008). *Classification of Industries.* Retrieved December 12, 2014, from https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaft klassifikationen/Content75/KlassifikationWZ08.html.

Federal Bureau of Statistics (2014). *Die zehn größten Unternehmensinsolvenzen in Deutschland im Jahr 2013.* Retrieved September 11, 2014, from http://de.statista.com/statistik/daten/studie/12997/umfrage/groesste-unternehmensinsolvenzen-in-deutschland/.

Fücks, R. (2013). *Intelligent wachsen.* München: Carl Hanser.

Glaser, B. G. (1978). *Theoretical Sensitivity.* San Francisco: The Sociology Press.

Helfrich, M. (2010). *Kreditscoring und Scorewertbildung der Schufa.* Münster.

Kaplan, A.M., and Haenlein, M. (2010). "Users of the world unite! The challenges and opportunities of Social Media". *Business Horizons,* 53 (1), 59-68.

Kramer, J.W., and Nitsch, K.W. (2010). *Facetten der Unternehmensfinanzierung.* Bremen: CT Salzwasser-Verlag GmbH.

Neuendorf, K.A. (2002). *The Content Analysis Guidebook.* Chicago: Sage Publications.

Schumann, M. (2002). "Bonitätsbeurteilungen und Kreditprüfungen im Electronic Business". *Electronic Business: theoretische Aspekte und Anwendungen in der betrieblichen Praxis*, 413-429.

Stiglitz, J.E. (1988). *Principal and Agent*. Department of Economics/Woodrow Wilson School of Public and International Affairs, Princeton University.

62. ESTIMATING CUMULATED ABSORBED DOSES AND ASSOCIATED HEALTH RISKS DUE TO OCCUPATIONAL EXPOSURE TO IONISING RADIATION

D. Moriña, E. Pernot and E. Cardis
Centre for Research in Environmental Epidemiology (CREAL), Universitat Pompeu Fabra (UPF), CIBER Epidemiología y Salud Pública (CIBERESP), Spain, dmorina@creal.cat

J. Grellier
Department of Epidemiology and Biostatistics, Imperial College, UK

A. Carnicer
Institute of Radiation Physics (IRA), Department of Radiology, Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland

ABSTRACT

Those working in interventional cardiology (IC) and related medical procedures are potentially subject to considerable exposure to X-rays to the eye. The lens of the eye is among the most radiosensitive tissues in the human body: ocular radiation exposure results in characteristic lens changes that, with time, may progress to partial or total lens opacification (cataracts). The severity of such changes tends to increase progressively with dose and time until vision is impaired and cataract surgery is required. Scattered radiation doses to the eye lens of an interventional cardiologist in typical working conditions can exceed 34 μGy/min in high-dose fluoroscopy modes and 3 μGy per image during image acquisition (instantaneous rate values) when radiation protection tools are not used. These are among the highest occupational radiation doses documented for medical staff whose work involves exposures to X-rays. We present a tool featuring an easy-to-use web interface that provides a general estimation of both cumulated absorbed doses experienced by medical staff

exposed in the interventional cardiology setting and their associated health risks. The tool is already available at http://crealradiation.shinyapps.io/radtool

1. INTRODUCTION

Interventional cardiology comprises a variety of minimally-invasive procedures used in the diagnosis and treatment of cardiovascular disease, and interventional radiology (IR) forms a key component of this work. In IR, hard X-rays (photon energies typically above 5-10keV) are used to create real time imaging of a patient undergoing IC surgery. Since interventional cardiologists and electrophysiologists carry out their work in close proximity to the patient on whom the imaging is being carried out, they are exposed occupationally to ionising radiation under normal working conditions.

Used appropriately to support a variety of procedures, IR provides enormous clinical benefits over other surgical procedures, including minimal invasiveness, reduced pain and risk of complications, shorter hospital stays and lower costs (Thom *et al.*, 2006; Togni *et al.*, 2004). The benefits of catheterisation over open surgery have resulted in a considerable increase in workloads over the past two decades and, although concomitant improvements in technology and radiation protection have reduced doses per procedure, there is concern that present cumulated doses may result in increased risk to surgeons working in IC of cataracts, as well as brain tumours or other neoplastic diseases. Effective use of radiation protection measures can lower the magnitude of associated health risks by reducing doses to the potentially exposed organs. In addition to the widely-used lead aprons and thyroid shields, a number of additional protection measures may be employed during IC procedures to reduce operator doses. These include, for example, flexible blanket-type shields laid over the patient during interventions to reduce operator exposure from scattered radiation, table skirts (screen suspended between the operating table and the floor), radiation protection cabins (several screens assembled so as to surround the operator, with apertures for the hands), ceiling-suspended screens (maneuverable screens that can be pulled down in front of the operator's face), lead eye glasses and protective surgical caps. The degree to which such measures are employed vary by procedure, specialisation, individual, hospital and country. Internationally, guidelines recommend that physicians involved in interventional procedures wear a leaded apron and a thyroid guard as a bare minimum radiation

protection measure. As such, the radiosensitive organs that are chiefly exposed are those in the head, in particular the eyes and the brain.

Radiation-induced cataract has been recognised as a highly relevant non-cancer endpoint among those exposed to ionising radiation since the 1930s (Merriam and Worgul, 2014). Radiation-induced cataracts typically develop as central opacities in the subcapsular posterior region of the lens, and consist of small granules and vacuoles that form a roughly circular opacity. Defects in lens transparency cause little or no visual impairment in the early stages of the disease, but eventually cause distortion and clouding. The reaction of the lens to radiation is partly attributable to the lifelong continual differentiation of the epithelial cells that make up the lens (fibregenesis); aberrant differentiation of cells due to exposure to radiation results in their accretion to the superficial posterior cortex (Worgul *et al.*, 2007), which forms a posterior subcapsular cataract (PSC).

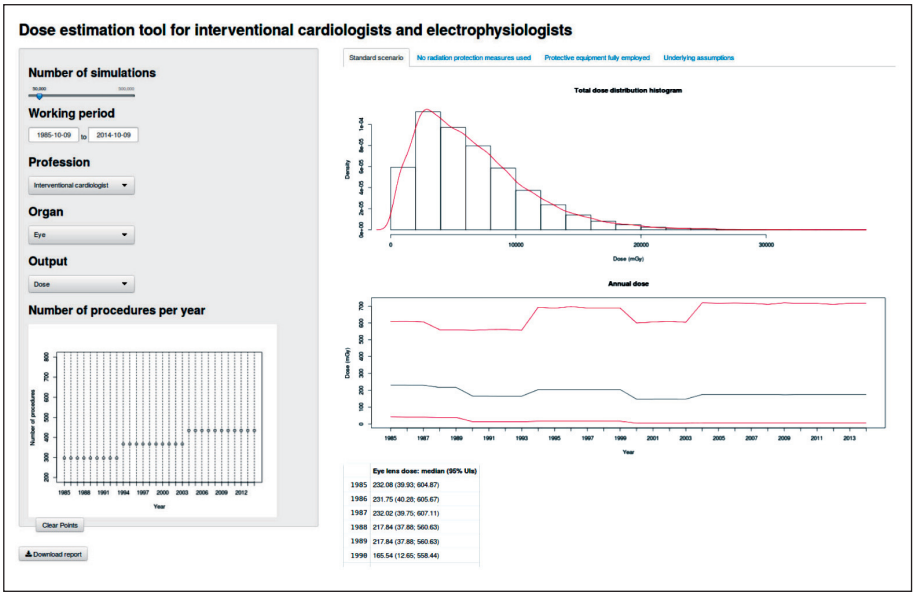


Figure 1. Screenshot of the web-based tool.

Accurate estimation of health impacts to medical staff from ionising radiation under a variety of operational and radiation protection scenarios is increasingly important as IC procedures are more frequently used. Of the existing online tools for radiation risk

assessment (e.g. Berrington *et al.*, 2012), the majority are focused on cancer effects (see Berrington *et al.*, 2012; or Hanna and Brewer, 2014, for instance). Our primary objective was to estimate doses and associated health impacts due to chronic occupational exposures typical in IC, including the risk of cataracts. To meet this objective, we aimed to design a tool that produces distributions of annual and total cumulated absorbed doses to the eye lens by employing robust estimators of parameters based on a multiple linear regression of predictors of dose.

2. METHODS FOR DOSE ESTIMATION AND RISK ASSESSMENT

The webtool has been developed in R (R Core Team, 2014) using the shiny package (RStudio Inc, 2014), which allows construction of interactive web applications from R. The user interface (Figure 1) comprises two panels, side by side. In the left-hand panel, the user enters specific data concerning their career (profession, work period and annual numbers of procedures), and the required output in terms of either dose or population attributable risk for a given population of health professionals for whom group-level occupational histories have been entered. Results are automatically generated and presented in the right-hand panel.

The output panel is divided into four tabs, the first three allowing the user to easily compare the dose or risk associated with a "standard" career, those expected when using all radiation protection measures available in that time period, and those expected when using no protection whatsoever. The fourth tab shows a summary of the most common procedures and usage of protection methods per decade according to the literature.

The potential predictors of dose included the usage of radiation protection measures, catheterisation access route, tube configuration and operator experience (derived from data collected in the ORAMED (Optimization of RAdiation protection for MEDical staff) project and from the literature) and a user-defined occupational history. We sought to take potential sources of uncertainty into account by means of Monte Carlo simulation, which allows uncertainties in model inputs to be propagated through to results. Occupational histories can be reconstructed by the user through general data derived from existing databases on typical numbers of

specific procedures carried out, radiation protection measures used, and X-ray tube configuration. Changes in imaging equipment typically available to interventional cardiologists in Europe over the past four decades were taken into account by fitting a metaregression model using results from a literature review, and the computed doses are adjusted using these values. Probability distributions of relative risk were calculated on the basis of the resulting cumulated absorbed doses, using estimates of dose-response and related uncertainties derived from the literature. In direct support of radiation protection, the tool allows the user to compare their estimated doses and the risks associated with them with those expected under a scenario where protective equipment is employed to the maximum possible extent.

Outputs of the tool allow population attributable risks of both health outcomes to be calculated for specific populations of these health professionals, for whom group-level occupational histories are reconstructed. By extension, it allows for estimation of the expected health benefits for that population associated with use of a variety of radiation protection measures.

In addition to the web version of the results, the user can download a report including all input and output data rendered as a pdf document.

The eye lens doses were estimated by means of a robust linear regression model that used data on usage of RadPad, table skirt, cabin and ceiling screen, the type of procedure, the tube configuration and operator experience (defined as high after 4 years of working in interventional cardiology or electrophysiology) as predictors of the absorbed dose. The obtained estimates are shown in Table 1.

The reduction of dose to the eye lens due to the usage of lead glasses was assumed to follow a PERT distribution (Vose, 2008) with minimal, modal and maximal values of 1, 3 and 10 respectively, based on expert opinion.

In addition to providing estimates of dose to the eye lens, the tool also provides the user with estimates of the health impacts associated with cumulated doses of ionising radiation, in terms of risk of radiation-related cataracts. The tool was constructed to offer several impact metrics: at the individual level, impacts are provided in terms of percentage increased relative risk of cataracts, and at the population level, in

terms of the population attributable risk, *AFp*. This risk estimation calculation required a dose-response function to be derived from the literature and we used odds ratio estimates at 1Gy from the literature (Worgul *et al*., 2007) for three cataract endpoints (Table 2) and an assumption of linearity on the log odds scale to generate a probabilistic estimate of the regression coefficient (β) and its standard error (SE). We used dose-response data taken at 1Gy and therefore the estimated size of the effect at lower doses is just an linear simplification.

Although it's not a likely situation within the purpose of this tool (where we expect to deal with moderate doses), dose thresholds were taken into account for cataract risk estimation according to (Worgul *et al*., 2007).

Protection method	$\hat{\beta}$ (95% CI)
Table	0.746 (0.543; 1.026)
RadPad	0.828 (0.485; 1.415)
Screen	0.826 (0.604; 1.130)
Cabin	0.168 (0.055; 0.518)
Procedure	$\hat{\beta}$ (95% CI)
CA PTCA	Ref.
DSA PTA C	0.392 (0.193; 0.794)
DSA PTA LL	0.920 (0.612; 1.384)
DSA PTA R	0.600 (0.392; 0.906)
Embolisation	0.778 (0.565; 1.070)
ERCP	2.166 (1.445; 3.246)
PM/ICD	1.827 (1.265; 2.639)
RF ablation	0.965 (0.620; 1.500)
Tube configuration	$\hat{\beta}$ (95% CI)
Above	Ref.
Below	0.244 (0.164; 0.363)
Biplane	0.230 (0.142; 0.372)
Experience	$\hat{\beta}$ (95% CI)
High	Ref.
Low	1.063 (0.836; 1.351)

Table 1. Parameter estimates and confidence intervals.

Cataract type	OR	Lower 95% CI	Upper 95% CI
Stage 1-5	1.70	1.22	2.38
Early PSC	1.89	1.25	2.84
Stage 1 PSC	1.42	1.01	2.00

Table 2. Odds ratios (OR) with 95% confidence intervals (CI) for three types of cataract.

Cataract type	Threshold dose
Stage 1-5	0.50 Gy
Early PSC	0.50 Gy
Stage 1 PSC	0.35 Gy

Table 3. Threshold doses for cataracts.

The concrete thresholds considered for each type of cataract are detailed in Table 3.

ACKNOWLEDGMENTS

This work was financially supported by funding from the Swiss Federal Office of Public Health.

REFERENCES

Berrington, A., Apostoaei, I., Veiga, L.H.S., Rajaraman, P., Thomas, B.A., Owen, H., Gilbert, E., and Land, C. (2012). "RadRAT: A Radiation Risk Assessment Tool for Lifetime Cancer Risk Projection". *Journal of Radiological Protection*, 32 (3), 205-222.

Hanna, Z., and Brewer, C. (2014). "An Online Patient-Oriented Radiation Risk Assessment Tool to Project Cancer Risk Following Exposure to LOW ŞIONIZING RADIATION in Canada". *Value in Health*, 17 (3).

Merriam, G.R., and Worgul, B.V. (2014). "Experimental Radiation Cataract – Its Clinical Relevance". *Bulletin of the New York Academy of Medicine*, 4, 372-392.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

RStudio, Inc. (2014). *Shiny:* Web Application Framework for R.

Thom, T., Haase, N., Rosamond, W., Howard, V.J., Rumsfeld, J., Manoilio, T., Zheng, Z., Flegal, K., O'Donnell, C., Kittner, S., Lloyd-Jones, D., Goff, D.C., Hong, Y., Adams, R., Friday, G., Furie, K., Gorelick, P., Kissela, B., Marler, J., Meigs, J., Roger, V., Sidney, S., Sorlie, P., Steinberger, J., Wasserthiel-Smoller, S., Wilson, M., and Wolf, P. (2006). "Heart Disease and Stroke Statistics–2006 Update: A Report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee". *Circulation*, 6.

Togni, M., Balmer, F., Pfiffner, D., Mayer, W., Zeiher, A.M., and Meier, B. (2004). "Percutaneous Coronary Interventions in Europe 1992-2001". *European Heart Journal*, 25 (14).

Vose, D. (2008). *Risk Analysis – A Quantitative Guide*. New York: John Wiley & Sons.

Worgul, B.V., Kundiyev, Y.I., Sergiyenko, N.M., Chumak, V.V., Vitte, P.M., Medvedovsky, C., Bakhanova, E.V., Junk, A.K., Kyrychenko, O.Y., Musijachenko, N.V., Shylo, S.A., Vitte, O.P., Xu, S., Xue, X., and Shore, R.E. (2007). *Radiation Research*, 167 (2), 233-243.

63. EXTREME VALUE ANALYSIS OF RIVER LEVELS IN A HYDROMETRIC STATION IN THE NORTH OF PORTUGAL

M.M. Neves
Instituto Superior de Agronomia and CEAUL, Universidade de Lisboa, Portugal,
manela@isa.ulisboa.pt

H. Penalva
Escola de Ciências Empresariais do Instituto Politécnico
de Setúbal and CEAUL, Portugal,
helena.penalva@esce.ips.pt

S. Nunes
Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal,
CEAUL and CMA/FCT/UNL, Portugal,
sandra.nunes@esce.ips.pt

ABSTRACT

Modelling extreme events of a random sample has become more and more necessary and exigent. Many important areas of application such as finance, insurance, risk theory, hydrology, ocean engineering, telecommunications and environmental risk assessment deal with extreme events. Extreme value theory (EVT) is the branch of probability and statistics dedicated to characterising the behaviour of the extreme observations. The analysis requires tools that must be simple to use but need to consider complex statistical models in order to produce valid inferences. Using accurate, friendly, free and open-source software, such as the R environment, is of particular importance. In this work we will present a review of parametric and semiparametric frameworks in EVT. Daily mean river levels from the hydrometric station at Fragas da Torre, during the years from 1946/47 to 1996/97, are considered for illustrating R procedures and functions in extreme value analysis done under those two frameworks.

1. A BRIEF INTRODUCTION

Let $X_1, X_2, , \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with common unknown distribution function F. The asymptotic theory of sample extremes distribution has been developed based on analogue arguments to those of the central limit theory. First results date back to Fréchet (1927), Fisher and Tippett (1928) and Von Mises (1936), but were Gnedenko (1943) and De Haan (1970) who gave conditions for the existence of sequences $\{a_n\} \in \mathbb{R}+$ and $\{b_n\} \in \mathbb{R}$ such that the partial maxima, $M_n \equiv \max(X_1, X_2, , \ldots, X_n)$, suitably normalized, had a non-degenerate limiting distribution, i.e.,

$$\lim_{n\to\infty} P\{(M_n - b_n)/a_n \leq x\} = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \quad \forall x \in \mathbb{R}. \qquad (1)$$

G is a non-degenerate distribution function (d.f), called Extreme Value distribution and is usually denoted by EV. This function is given by

$$EV_\xi(x) = \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0,\ if\ \xi \neq 0\ ; \\ \exp[-\exp(-x)], & x \in \mathbb{R},\ if\ \xi = 0, \end{cases}$$

where the shape parameter ξ is called extreme value index, a primary parameter in Statistics of Extremes measuring the heaviness of the right-tail, $\overline{F} := 1 - F$.

For a function F verifying (1) we say that it is in the domain of attraction (for maxima) of EV_ξ and write $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$.

The $EV\xi$ incorporates the three (Fisher-Tippett) types and is the general form for the following distributions for maxima:

- Gumbel: $\Lambda(x) = \exp(-\exp(-x)) = EV_0(x), x \in \mathbb{R}, \xi = 0;$; it is the limit for exponential tailed distributions and the right endpoint can be either finite or infinite;
- Fréchet: $\Phi_\xi(x) = \exp(-x^{-1/\xi}) = EV_\xi(\frac{x-1}{\xi}), x > 0, \xi >$; it is the limit for heavy tailed distributions and the right endpoint is infinite;
- Weibull: $\Psi_\xi(x) = \exp(-(-x)^{1/\xi}) = EV_\xi(\frac{-x-1}{\xi}), x < 0, \xi < 0,$, it is the limit for short tailed distributions and the right endpoint is finite.

The EV_ξ d.f. can also incorporate location, λ, and scale, $\delta > 0$, parameters. The estimation of ξ was initially performed under a parametric framework. Gumbel's method or block maxima (BM), the method of Largest Observations (LO), the Multidimensional EV models and the Peaks Over Threshold (POT) method are the most well known methods in this approach. In the seventies semi-parametric approaches appeared, based on De Haan (1970) work. Here estimates are usually based on the largest $\boldsymbol{k}$ order statistics in the sample, or in the excesses over a high random threshold, assuming only that the model F underlying the data is in $D_M(EV_\xi)$ or in specific subdomains of $D_M(EV_\xi)$.

This work intends to review and comment statistical issues arising in modelling extremes of a random sample and doing statistical inference under both parametric and semi-parametric frameworks, using or constructing functions in R environment. A case study of daily mean river levels from the hydrometric station at Fragas da Torre, in the North of Portugal, will illustrate the procedures reviewed and the challenges that appear.

2. THE R SOFTWARE IN EXTREME VALUE THEORY

The software R (R Development Core Team, 2012) is an open source language for statistical computing, which allows the data analysis, numerical computation and inference in a wide range of statistical problems.

Several R-packages exist devoted to extreme value analysis, such as `evd`, `ismev`, `evir`, `extremes`, `fExtremes`, `extremevalues` `evdbayes`, `copula`, `POT`.

An excellent software review for extreme value analysis in several packages, can be seen in Gilleland *et al*. (2013). These authors gave a summary of available utilities in the packages analyzed.

Here we will consider a few packages in R and functions therein that will be explained and used in the study of our application. Some functions, for the recent estimators and methodologies needed to be constructed and included in our programmes.

3. EXTREME DATA ANALYSIS

The data set refers do daily mean river levels from hydrometric station at Fragas da Torre, during the years from 1946/47 to 1995/96. The 50 years observed correspond exactly to the period between October 1st of 1946 to September 30th of 1996. The source of Paiva river is in the Serra de Leomil, in the North of Portugal, and its hydrographic basin has an area of approximately 700 Km^2. The flow study of this river is a matter of major importance since it is one of the main alternatives to the Douro river as source of water supply in the south of Oporto region. In particular there is a project to build a dam precisely at Fragas da Torre.

In this study only the months from November until April were considered. Stationarity was verified. The basic statistics and the graphical display (chronogram, autocorrelation function, histogram and boxplot) of the data are shown in Table 1 and Figure 1, respectively.

min	1st Qu.	Median	Mean	3rd Qu.	max
0.00	9.07	17.5	35.14	38.10	920.0
	n	Skewness	Kurtosis	sd dev	
	9050	4.12	27.60	51.25	

Table 1. Basic statistics for daily mean river levels.

Skewness and Kurtosis measures as well as the graphical display reveal a tail heavier than the Gaussian one.

In the parametric approach several models have been considered. As an illustration let us show results obtained through the Block Maxima Method . The n-sized sample is divided into m sub-samples of size r (usually associated to m years, with $n = r \times m$, r reasonably large). In each of the m sub-samples the maximum of the r observations is considered.

From `evd` package and using `fgev()` and `confint()` functions the estimates and the confidence intervals (in this case we show 95%) were obtained

$\hat{\lambda}$	$\hat{\delta}$	$\hat{\xi}$
227.351	122.727	-0.018
(189.440, 265.263)	(95.542, 149.911)	(-0.197, 0.161)

Peaks Over Threshold, *Largest Observations* and *Multi-dimensional* EV models have also been used.

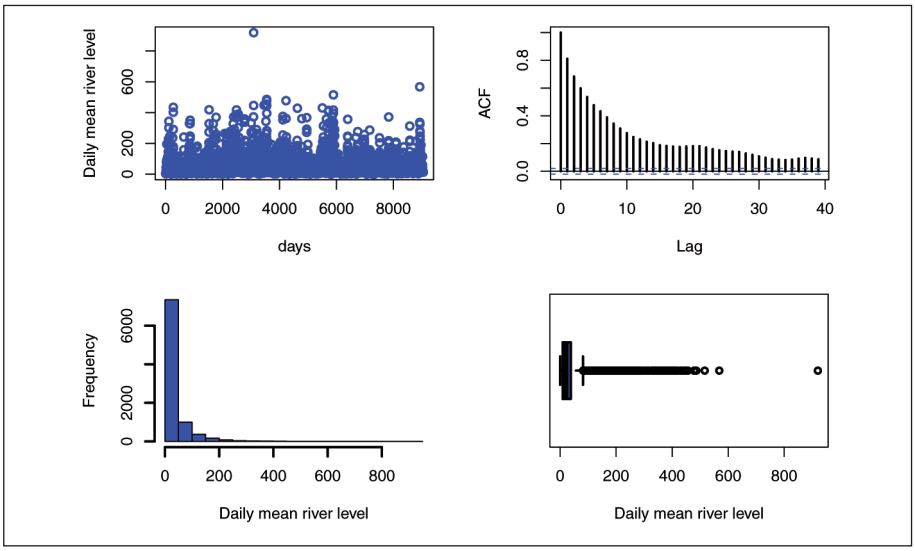


Figure 1. Chronogram, autocorrelation function, histogram and boxplot for daily mean river levels.

In the **semi-parametric approach** the reliable estimation of ξ is crucial. Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ be the ordered statistics associated to the sample $(X_1, X_2, , \ldots , X_n)$. The estimation is now performed on the basis of the k top observations, with k intermediate, i.e. such that $k = k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$. Many estimators exist in the literature. Here we will refer to the classical ones, such as those presented in Hill (1975), Pickands (1975) and Dekkers *et al.* (1989), respectively, and named as:

- Hill: $\widehat{\xi}^{H}_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \ln(X_{n-i+1:n}) - \ln(X_{n-k:n})$ (for $\xi > 0$)
- Pickands: $\widehat{\xi}^{P}_{k,n} := \frac{1}{\ln 2} \ln\left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}}\right)$ (for $\xi \in \mathbb{R}$)
- Moments: $\widehat{\xi}^{M}_{k,n} := M^{(1)}_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{(M^{(1)}_{k,n})^2}{M^{(2)}_{k,n}}\right)^{-1}$, (for $\xi \in \mathbb{R}$) with $M^{(r)}_{k,n} := \frac{1}{k}\sum_{i=1}^{k}[\ln(X_{n-i+1:n}) - \ln(X_{n-k:n})]^r$,

and also to more recent, reduced-bias estimators, such as the minimum-variance reduced-bias (MVRB) estimators, Caeiro, Gomes and Pestana (2005).

$$\widehat{\xi}^{H}_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \ln(X_{n-i+1:n}) - \ln(X_{n-k:n}) \text{ (for } \xi > 0)$$

where $\hat{\beta}$ e $\hat{\rho}$ are adequate consistent estimators of second-ordered parameters β and ρ, that appear in the second order condition required for obtaining higher developments of tail parameters' estimators distribution. These estimators have an asymptotic variance equal to the one of the Hill estimator but an asymptotic bias of smaller order, overpassing the classical ones for all k. Figure 2 shows the sample paths for the estimates obtained by those estimators.

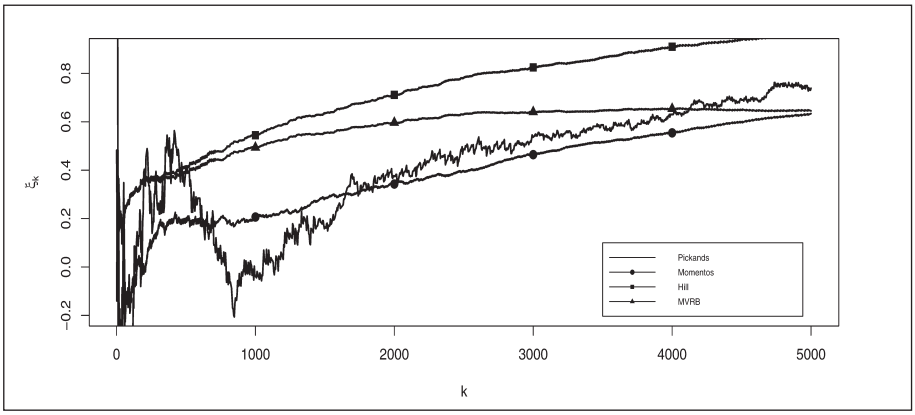


Figure 2. Sample path from Hill, Moments, Pickands and MVRB estimates.

The choice of the level k still remains an open research topic. Some suggestions have appeared such as the heuristic algorithm, Gomes, Martins and Neves (2013), with an adaptive choice of k for estimating ξ. This algorithm was also applied to our data set. Only as an illustration and given that the MVRB estimates show the more stable path, for these estimates that algorithm led to the following values: $\hat{k}$ = 3827 and $\hat{\xi}$ = 0.4502.

4. SOME COMMENTS AND REMARKS

This work follows previous works from the authors, Penalva *et al.* (2013, 2014) where R procedures for the analysis of extreme values were exploited. Comparisons, comments and challenges regarding estimation performed through parametric and semi-parametric approaches are discussed. Actually different values were obtained for ξ estimates through BM method and semi-parametric estimators. Looking for reasons for those different values is also the challenge of this work.

ACKNOWLEDGMENTS

Research partially supported by National Funds through FCT–Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014 (CEAUL)

REFERENCES

Caeiro, F., Gomes, M.I., and Pestana, D.D. (2005). "Direct Reduction of Bias of the Classical Hill Estimator". *Revstat* 3: 111-136.

Dekkers, A.L.M., Einmahl, J.H.J., and De Haan, L. (1989). "A Moment Estimator for the Index of an Extreme-Value Distribution". *Annals of Statistics* 17 (4), 1833-1855.

Fisher, R.A., and Tippett, L.H.C. (1928). "On the estimation of the frequency distributions of the largest or smallest member of a sample". *Proceedings of the Cambridge Philosophical Society,* 24, 180-190.

Fréchet, M. (1927). "Sur la loi de probabilité de l'écart maximum". *Ann. Soc. Polon. Math.* (Cracovie), 6, 93-116.

Gomes, M.I., Martins, M.J., and Neves, M.M. (2013). "Generalized Jackknife-Based Estimators for Univariate Extreme-Value Modeling". *Communications in Statistics, Theory and Methods*., vol. 42, Issue 7, 1227-1245.

Gilleland, E., Ribatet, M., and Stephenson, A.G. (2013). "A Software Review for Extreme Value Analysis". *Extremes*, 16, Issue 1, 103-119.

Gnedenko, B.V. (1943). "Sur la distribution limite d'une série aléatoire". *Annals of Mathematics*, 44, 423-453.

De Haan, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract 32, Amsterdam, Dordrecht: D. Reidel.

Hill, B.M. (1975). "A Simple General Approach to Inference about the Tail of a Distribution". *Annals of Statistics* 3, 1163-1174.

Penalva, H., Neves, M.M., and Nunes, S. (2013). "Topics in Data Analysis Using R in Extreme Value Theory". *Advances in Methodology and Statistics / Metodoloski zvezki* 10 (1), 17-29.

Penalva, H., Nunes, S., and Neves, M.M. (2014). "Statistical Modeling and Inference in Extremes: Applications with R.". *Biometrie und Medizinische Informatik Greifswalder Seminarberichte, Statistical and Biometrical Challenges*, Shaker Verlag, 281-309.

Pickands, J. (1975). "Statistical Inference Using Extreme Order Statistics". *Annals of Statistics 3,* 119-131.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R. Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. ISBN 3-900051-07-0.

Von Mises, R. (1936): "La distribution de la plus grande de n valeurs". Reprinted in *Selected Papers*, Volumen II, American Mathematical Society, Providence, R.I. (1954), 271-294.

64. ON MEASURES OF FINANCIAL RISK

S. Novak
Middlesex University London, School of Science and Technology, UK,
and Tomsk State University, Russia,
S.Novak<at>mdx.ac.uk

ABSTRACT

Traditional measures of financial risk (e.g., the standard deviation, Value-at-Risk (VaR), conditional Value-at-Risk (CVaR)) are widely used by banks when deciding on the amount of capital they need to set aside in order to offset the market risk. However, the traditional risk measures appear static; they barely change with the inflow of new data. Dynamic risk measures look capable of indicating increased risk on the eve of a sharp market movement. However, they swing too widely to be used when deciding on the amount of the capital reserve.

We argue that a combination of static and dynamic risk measures may serve the purpose, staying conservative in normal situations and suggesting the increase of the amount of the capital reserve during volatile periods.

1. INTRODUCTION

The problem of accurate evaluation of financial risk is fundamental to financial well-being of individual investors as well as financial institutions. However, the issue is demanding and requires sophisticated statistical analysis of financial data.

Modern methods of risk measurement are capable of forecasting the magnitude of a possible sharp market movement (Novak & Beirlant, 2006; Novak, 2011). Yet the estimates appear static: they barely change with the inflow of new information, and

may be unsuitable for active portfolio management. A number of dynamic risk measures are available to investors and risk managers (see, e.g., Novak, 2011). However, they are still lacking proper statistical scrutiny.

In this paper we compare a number of risk measures. We argue that a combination of static and dynamic risk measures may be more suitable for the purpose of risk evaluation and forecasting than the traditional approach. In the next section we briefly overview popular measures of risk. Section 3 discusses methods of risk measurement from dependent heavy-tailed data. Section 4 concentrates on dynamic risk measures.

2. POPULAR MEASURES OF FINANCIAL RISK

Let X denote the quantity of interest (say, the rate of return of a particular financial asset, the daily log-return of a share price, etc). Historically, the first measures of risk were the standard deviation σ_{X_n} and the variance (Markowitz , 1952).

With the development of the CAPM model beta was added to the list of measures of risk. In certain models (e.g., ARCH/GARCH) one deals with the conditional standard deviation σ_{X_n} of the quantity of interest X_n given the previous observations $X_n{-}1$, $X_n{-}2$, …

Among all traditional measures of risk Value-at-Risk (VaR) seems to be the most popular one. Let

$$F(x) = \mathbb{P}(X \leq x)$$

be the distribution function (d.f.) of $\mathcal{L}(X)$. Up to a sign, Value-at-Risk is an extreme quantile:

$$m\%\text{-VaR} = -F^{-1}(m/100)\,.$$

Equivalently, q-VaR $= -F^{-1}(q)$, where $0 < q < 1$. Recall that

$$F^{-1}(q) = \inf\{x : F(x) \geq q\}\,.$$

If F is continuous, then q-VaR is such a number that

$$\mathbb{P}(X \leq -q\text{-VaR}) = q.$$

The value of $m\%$-VaR indicates how far the quantity of interest (say, daily log-return X) can fall in approximately m% "worst" cases. For instance, if 1%-VaR equals –0.04, then in approximately 1% cases the quantity of interest, X, can be below -4%.

Knowing VaR allows a bank to decide how much money it needs to put aside to offset the risk of an undesirable market movement. For instance, Goldman Sachs deals with 5%-VaR; Citigroup, Credit Suisse First Boston, Deutsche Bank, JP Morgan Chase and Morgan Stanley use 1%-VaR.

Closely related to VaR is the measure of risk called Conditional VaR (CVaR). It represents the average loss given there is a fall beyond VaR:

$$\text{CVaR} = -\mathbb{E}\{X | X \leq -\text{VaR}\}.$$

One may prefer dealing with positive numbers (for instance, one can speak about "20.5% fall" of the S&P500 index on the "Black Monday" instead of "–20.5% rate of return"). If we switch from X to $-X$, then VaR is defined as the inverse of $F_c = 1 - F$:

$$q\text{-VaR} = F_c^{-1}(q)\,, \tag{1}$$

and

$$\text{CVaR} = \mathbb{E}\{X | X \geq \text{VaR}\} \tag{2}$$

(we sometimes omit prefix q-). Properties of VaR and CVaR can be found, e.g., in Alexander (2008), Novak (2011), Pflug (2000).

Note that σ, VaR and CVaR evaluate risk on the base of a sample of the past ("historical") data.

If data is light-tailed, it hardly can exhibit extreme movements, and the assumption that $\mathcal{L}(X)$ has normal $\mathcal{N}(\mu; \sigma^2)$ distribution is not unreasonable.

In the case of frequent data parameter μ is typically negligent, and hence

$$q\text{-VaR}_{\mathcal{N}} \simeq -\sigma t_q\,,\ q\text{-CVaR}_{\mathcal{N}} \simeq \exp(-t_q^2/2)\sigma/\sqrt{2\pi}\,, \tag{3}$$

where $t_q = \Phi^{-1}(q)$ and Φ is the standard normal distribution function. According to (3), one only needs to estimate the standard deviation σ (cf. Alexander (2001), §9.3).

If $\{X_n, n \geq 1\}$ is the ARCH(1) process with parameters $b > 0$ and $c \geq 0$:

$$X_n = \xi_n \sqrt{b + cX_{n-1}^2} \qquad (n \geq 2)\,, \tag{4}$$

where $\{\xi_i\}$ is a sequence of independent normal $\mathcal{N}(0; 1)$ random variables, then the conditional standard deviation of X_n given (X_{n-1}, X_{n-2}, ...) is $\sigma_{X_n} = \sqrt{b + cX_{n-1}^2}$,

$$\begin{aligned} q\text{-VaR}_{X_n} &\simeq -t_q\sqrt{b + cX_{n-1}^2}\,, \\ q\text{-CVaR}_{X_n} &\simeq \exp(-t_q^2/2)\sqrt{b + cX_{n-1}^2}\,\Big/\sqrt{2\pi}\,. \end{aligned}$$

The estimate of parameter c is typically close to 0, making σ_{Xn}, q-VaR$_{Xn}$ and q-CVaR$_{Xn}$ rather static. Note that the stationary distribution of the ARCH(1) process is heavy-tailed; it is a "short-memory" process (Goldie, 1991; Embrechts *et al.*, 1997, pp. 465-466).

3. RISK ESTIMATION FROM HEAVY-TAILED DATA

Financial/insurance data often exhibits heavy tails (Mandelbrot, 1963; Fama and Roll, 1968; Embrechts *et al.*, 1997, p. 404). This is particularly common to "frequent" data (e.g., daily log-returns of stock prices and stock indexes), while log-returns of less frequent data can exhibit lighter tails.

The feature of heavy-tailed distributions is that a single observation can be of the same order of magnitude as the whole sum of sample elements: a single claim to an insurance company or a one-week market movement can cause a loss comparable to a one-year profit.

The problem of evaluating risk from dependent heavy-tailed data attracts increasing attention of researchers (see Embrechts *et al.*, 1997; Novak, 2011, and references therein). A number of procedures to check if the distribution is heavy-tailed are mentioned in Markovich (2007) and Novak (2011).

The distribution of a random variable X has a heavy right tail if

$$\mathbb{P}(X \geq x) = L(x)x^{-\alpha} \qquad (\alpha > 0), \tag{5}$$

where the (unknown) function L is slowly varying at ∞:

$$\lim_{x\to\infty} L(xt)/L(x) = 1 \qquad (\forall t > 0).$$

Number α in (5) is called the tail index. It is the main characteristic describing the tail of a heavy-tailed distribution.

The problem of reliable estimation of the tail index, $y_q = q$–VaR and $z_q = q$–CVaR from heavy-tailed data is demanding (see Embrechts *et al.*, 1997; Novak, 2011, and references therein).

The following estimators of VaR and CVaR are asymptotically normally distributed and appear reasonably accurate in examples of simulated data (see Novak, 2011):

$$y_{n,q} \equiv y_{n,q}(x) = (N_n(x)/qn)^{\hat{a}_n^{RE}} x\,, \tag{6}$$

$$z_{n,q} \equiv z_{n,q}(x) = y_{n,q}/(1-\hat{a}_n^{RE})\,, \tag{7}$$

where threshold x is a tuning parameter (it needs to be chosen), q is the given level and $\hat{a}_n^{RE}$ is the Ratio Estimator of index $a = 1/\alpha$:

$$\hat{a}_n^{RE} \equiv \hat{a}_n^{RE}(x) = \sum_{i=1}^{n} \ln(X_i/x)\mathbb{1}\{X_i > x\} \Big/ \sum_{i=1}^{n} \mathbb{1}\{X_i > x\} \tag{8}$$

(equivalently, $1/\hat{a}_n^{RE}$ is the Ratio Estimator of the tail index).

This approach is non-parametric. There are doubts that parametric models accurately describe real financial data: one usually cannot be sure if the unknown

distribution comes from a chosen parametric family. The advantage of the non-parametric approach is that such a problem is void: the non-parametric class is so rich that one typically has no doubt the unknown distribution belongs to it. The disadvantage of the non-parametric inference is the presence of a tuning ("nuisance") parameter.

The comparison of finite-sample properties and asymptotic mean-squared errors of a number of tail index and VaR estimators in Novak (2011) is in favor of the approach.

A procedure of choosing the tuning parameter x for estimators (6), (7) and (8) can be found in Novak (2011). Based on theoretical results, it is data-driven and aims to "let data speak for itself".

When tested on samples of simulated data, as well as applied to the problem of risk evaluation on the eve of the "Black Monday" crash, the method has demonstrated better accuracy than other approaches.

4. DYNAMIC MEASURES OF RISK

The measures of risk discussed in the previous sections are static, they barely change with the inflow of new observations. This is unavoidable as every such measure is a function of a sample of preceding observations, and the sample size should be large enough to make estimation accurate. Probably the first step towards dynamic risk measurement was the introduction of the ARCH process (4). The conditional standard deviation σ_{Xn} is a function of the previous observation. However, the change from σ_{Xn} to σ_{Xn+1} is typically small.

Truly dynamic is measure m_{TA} that incorporates ideas of loss reduction (see Novak, 2011). Measure m_{TA} does change considerably as price changes. We present properties of m_{TA} and then compare the behaviour of m_{TA} and static risk measures on examples of real data.

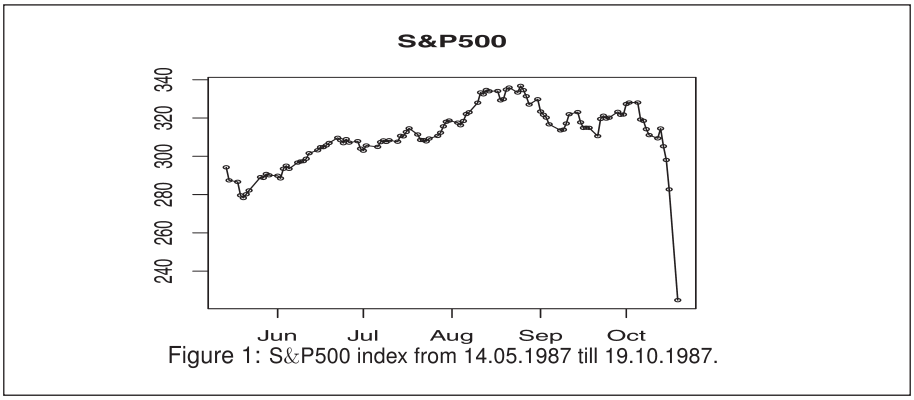


Figure 1. S&P500 index from 14.05.1987 till 19.10.1987.

Example. From mid-August 1987 the S&P500 index was mainly declining (see Fig. 1). Table 1 and Fig. 2 show steady decrease of the standard deviation of the S&P500 index since early September 1987 till 16 Oct 1987, suggesting lower level of risk. The behaviour of m_{TA} is very different: Fig. 2 shows spikes of m_{TA} on the eve of the Black Monday.

Date	S&P500	σ	m_{TA}
8.10.1987	314.16	28.55	3.62
9.10.1987	311.07	28.42	0.53
12.10.1987	309.39	28.27	6.45
13.10.1987	314.52	28.14	3.98
14.10.1987	305.23	27.99	2.29
15.10.1987	298.08	27.84	9.62
16.10.1987	282.7	27.67	4.49

Table 1. S&P500 index, its standard deviation and mTA on the eve of the Black Monday.

On 9.10.1987 the value of m_{TA} was 0.53. The next day the market has shown a 12-fold increase of the level of risk. The data was ringing a bell but would anyone hear it using the static measures?

5. CONCLUSION

The paper compares available measures of financial risk on the eve of a market downturn. We argue that a combination of static and dynamic risk Standard deviations of S&P500 measures may have an advantage over the use of a single measure, staying conservative in normal situations and prompting to increase of the amount of the capital reserve during volatile periods.

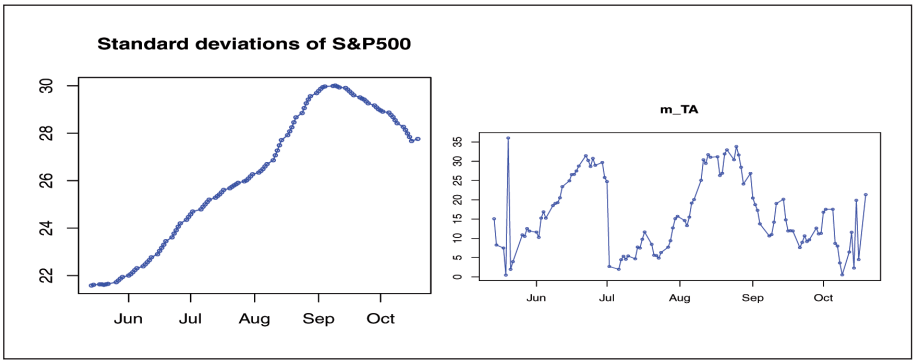


Figure 2. The standard deviations of the S&P500 index (left) and mTA (right) for the period 14.05.1987 – 19.10.1987.

REFERENCES

Alexander, C. (2001). *Market Models*. New York: Wiley.

Alexander C. (2008). *Value-at-Risk Models*. New York: Wiley.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

Fama, E.F., and Roll, R. (1968). "Some Properties of Symmetric Stable Distributions". *J. Amer. Statist. Assoc.*, 63, 817-836.

Goldie, C.M. (1991). "Implicit Renewal Theory and Tails of Solutions of Random Equations". *Ann. Appl. Probab.*, 1, 126-166.

Mandelbrot, B.B. (1963). "New Methods in Statistical Economics". *J. Political Economy*, 71, 421-440.

Markowitz, H.M. (1952). "Portfolio Selection". *J. Finance*, 7, No 2, 77-91.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data*. Chichester: Wiley.

Novak, S.Y., and Beirlant, J. (2006). "The Magnitude of a Market Crash Can Be Predicted". *J. Banking & Finance*, 30, 453-462.

Novak, S.Y. (2011). *Extreme Value Methods with Applications to Finance*. London: CRC. ISBN: 978-1-43983-574-6.

Pflug, G.C. (2000). "Some Remarks on the Value-at-Risk and the Conditional Value-at-Risk". *Probabilistic Constrained Optimization*, 272-281. Kluwer: Netherlands.

65. STOCHASTIC RESPONSE SURFACE METHODOLOGY: A TOOL TO RISK ASSESSMENT

T.A. Oliveira
Department of Sciences and Technology, Open University, CEAUL Portugal
teresa.oliveira@uab.pt

C. Leal
Department of Sciences and Technology, Open University, Portugal
conceicao.leal2010@gmail.com

A. Oliveira
Department of Sciences and Technology, Open University, CEAUL Portugal
amilcar.oliveira@uab.pt

ABSTRACT

The Response Surface Methodology (RSM) plays a key role on the optimization of Experimental Design problems, once response surface models allow to approximate the structural data behavior as accurate as possible. Following the review of RSM applications on risk assessment, presented in Oliveira *et al.* (2015), the approach to the stochastic form of RSM implementation is now explored. The analysis of Wisconsin Breast Cancer Prognostic database has been performed by comparing uncensored data with the right censored data. The models obtained with uncensored date suffer from the right bias and so the main aim of our work is to provide an illustration on how to attain more realistic results than those obtained only considering data with relapsed and to provide more consistent models for the prediction and risk analysis.

Computational results and graphics were obtained by using R software.

1. INTRODUCTION

Risk analysis is the process of systematically identifying and assessing potential risks and uncertainties, which occur in a more or less complex system and then to look for a viable strategy that allows a more efficient manage of these risks.

In probabilistic risk assessment, uncertainty is considered and the risk is characterized by a probability distribution. The respective model is then used to create / simulate different risk scenarios.

The numerical simulation used to generate data for analysis often involves high computational cost, such that it requires the use of metamodels. The RSM has proved to be a suitable tool to estimate metamodels, both for systems of behaviors and risk assessment or for the quantification of uncertainties, revealing in this case a good alternative to Monte Carlo simulation. The Monte Carlo simulation, associated with Stochastic Response Surface Methodology (SRSM) provides an effective way to study the propagation of uncertainty in the response. SRSM is a powerfull tool for modeling uncertainty in simulation models, mainly in the context of Risk Analysis. Developments on SRSM have been remarkable over the past few years, mainly due to the technological advances and modern equipments in what concerns hardware and software. It has been possible to extend and to explore this methodology in many areas and to state its importance in studies connected with Health problems. More specifically in Medicine, it became a crucial tool for specialists to attain better and accurate Diagnosis and Prognosis. In recent papers, see Leal *et al.* (2014) and Oliveira *et al.* (2015), the authors paid attention to the SRSM applications and behavior, in particular when it is applied to a sample with uncensored data.

2. STOCHASTIC RESPONSE SURFACE METHODOLOGY (SRSM) AND RISK

Monte Carlo method is the most popular method used on probabilistic quantification of covariate uncertainty propagation in the system response and on the estimation of its probability distribution. However, this may have very high computational costs and it urges to rely on methodologies that converge more quickly to the solution.

The SRSM allows the generation of a metamodel, computationally less demanding and statistically equivalent to the complete numerical model.

Isukapalli *et al*. (1998) introduced this methodology and presented two case studies. The basic idea of the methodology is to represent the response of a model to changes in covariate, using a response surface defined with a polynomial basis that is orthogonal with respect to a probability measure on the space of parameters.

In the classic version of the methodology, a vector of random variables $\xi = (\xi_i)$, $i = 1, \ldots, n$, is sellected, under $N(0,1)$ distribution, representing uncertain variables $x = (x_i)$ of the model, in such way that $x_i = h(\xi_i)$. After this selection, response variables are represented as a function of the same vector of random variables, $Y = f(c,\xi)$, being a vector c of coefficients to estimate. The response of the system model to the various achievements of x, permits the estimation of model coefficients. The coefficients c_i quantify the dependence of the response Y on the input vector x, for each realization of x. The form of the function f result of the polynomials chaos expansion Ψ_i polynomials which form a base of orthogonal polynomials to a given probability measure) and is expressed by:

$$Y = f(c,\xi) = c_0 Y_0 + \sum_{i1=1}^{\infty} c_{i1}\Psi_1(\xi_{i1}) + \sum_{i1=1}^{\infty}\sum_{i2=1}^{i1} c_{i1i2}\Psi_2(\xi_{i1},\xi_{i2}) + \cdots$$

This series is approximate by a truncated polynomial in *n-th* power. The roots of the polynomial of $n + 1$ degree are used in the simulation model to get the data to estimate the model coefficients.

The methodology is implemented sequentially as follows: (i) representation of uncertain input variables; (ii) representation of the response variable; (iii) estimation of the model parameters; (iv) obtain the response statistical properties; (v) evaluation of the model approximation.

In the case of the classical approach, the measure is the Gaussian and the polynomials are the Hermite polynomials which form is as follows:

$$PCE = c_0 + \sum_{i=1}^{n} c_i\xi_i + \sum_{ii=1}^{n} c_{ii}(\xi_i^2 - 1) + \sum_{i=1}^{n-1}\sum_{j>i}^{n} c_{ij}\xi_i\xi_j$$

Oladyshkin and Nowak (2012) proposed a new generalization of the methodology, data-driven chaos expansion, in which statistical moments are the only source of information that is propagated in the stochastic model. Probability distributions may be discrete, continuous, or continuous discretized, may be specified by analytical way (through PDF or CFD), numerically using a histogram or by using the raw data. In this approach, all distributions are admissible for the input variables of a given model, as long as they have in common a finite number of moments. Thus, in the case of considering truncated polynomial, there is just the need to know a finite number of moments, with no need for complete knowledge of the probability density function or even its existence, which frees the researcher from the need of assumptions which may not always be supported by the existing data. According to the literature it is known that this expansion converges exponentially and faster than the classical expansion.

The estimation of model parameters depends on the model complexity, Isukapalli *et al.* (1998). In the case of invertible model, the parameters can be obtained directly from the input random variables $(\xi_i)_{i=1}^{n}$. If the model equations are mathematically manipulated, in spite of nonlinearities, then the model coefficients can be obtained afterwards, by an appropriate norm minimization of residuals, replacing the input random variables by the respective transformations in terms of Gaussian variables $N(0,1)$. Isukapalli *et al.* (1998) present and discuss some methods for parameter estimation, all based on the collocation points methods: Probabilistic Collocation Method, Efficient Collocation Method and Regression Based Method.

The expansion in polynomials chaos is a simple but powerful tool for stochastic modeling. Probability density function, probability distribution functions or other statistics of interest can be estimated and quickly evaluated via Monte Carlo simulation, once the evaluation of a polynomial function is faster than the original equations model evaluation.

In case of using data-driven expansion for Risk Analysis, one can directly use a set of large-sized data or a probability density function of maximum or relative minimum entropy, since, in this case, the relevant moments of the polynomial chaos expansion are compatible with those of the input variables. The bootstrap resampling method may be used to obtain more precise estimates of the moments from a

reduced set of available variables, providing a more accurate estimation of the risk assessment model.

3. APPLICATIONS TO RISK ASSESSMENT

The RSM has an important role in generating reduced models to replacing the simulator complex processes that require a very high number of simulations. Applications are diverse and many of them relate to the SRSM for quantification of uncertainty in stochastic processes.

Oladyshkin and Nowak (2012) applied the SRSM, based on data-driven polynomial chaos expansion, to a problem of contaminant transport in 3D heterogeneous aquifer and the risk to human health caused by exposure of the population. Datta (2013) applied the SRSM to estimate the propagation of uncertainties in the parameters of the function retention of strontium in the human body. The uncertainty parameters, the heterogeneity of the patient and the stochastic uncertainty results are increasingly important concepts in models of medical decision. Dittus *et al.* (1989) and Koerkamp *et al.* (2010) show various methods for analyzing uncertainty and the heterogeneity of patients in decision models.

The breast cancer prognosis is markedly heterogeneous, and research has often focused on the effect of prognostic factors related to the disease, such as the expression of estrogen receptors, tumor size and others. Still remains an open question of modeling the data of the time to recur, the complexity of the shape of the hazard function over follow-up period, and the identification of factors that may affect it, using a fully parametric approach (Ardoino *et al.*, 2012).

The application of RSM in association with other techniques such as neural networks, computer simulation or genetic algorithms can be found in many applications going from Industries, Physical and Chemical Sciences, Engineering, Biological and Medical Sciences, Food Sciences, Social Sciences, Agriculture, Aeronautics to other (and countless) areas. Response surface approximations serve as surrogate models for the full mathematical model that can be used to quickly explore regions of the input space that were not sampled. The risk assessment is a field where the

RSM can be used, namely for modeling consequences from an event and the uncertainties in the model, for which the SRSM is very useful.

4. APPLICATION IN HUMAN HEALTH

In our work SRSM was applied to the Wisconsin Breast Cancer Prognostic (WPBC) database, constituted by data of 253 patients with breast cancer, who have undergone surgical excision of invasive cancer. First, only 69 of these cases were used, namely those that correspond to the patients who had remission till the end of the study. Second, all cases were used, including 184 that are right censored (patients that withdrew from the study or for those for which only the time of the last examination was known). The database is very well described by Wolberg *et al.* (1999), who are responsible for its design and construction. Using regression on sample data, a first and a second order polynomial chaos expansion were estimated with Hermite polynomials, in which the response variable is the time to recur (TTR) and the uncertainty parameters are two or three of the ten available covariates, SIZE-related with the tumor characteristics, TEXTURE and AREA - related to the morphological characteristics of malignant cells nucleus.

Different models were simulated with respect to the polynomial degree and to the number and nature of the covariates. The second order models, with two covariate, were used to estimate the combination of values that minimize the TTR and offer the patient the worst scenario. The second order model, with three covariate, was used to estimate the distribution of the response variable and the resulting survival and risk functions. The survival functions (disease-free survival time) were estimated by $S(t) = 1 - F(t)$, where F is the distribution function of TTR, and were compared with the Cox model.

To estimate the hazard function, which is defined by $H(t) = \frac{f(t)}{S(t)}$, $f(t)$ being the density function of TTR, there is the need to get the theoretical distributions which best fit $S(t)$ the survival functions. These were Gamma and Lognormal functions to uncensored and censored data, respectively, so it was possible to estimate H and to obtain its graphical representations (Figure 1).

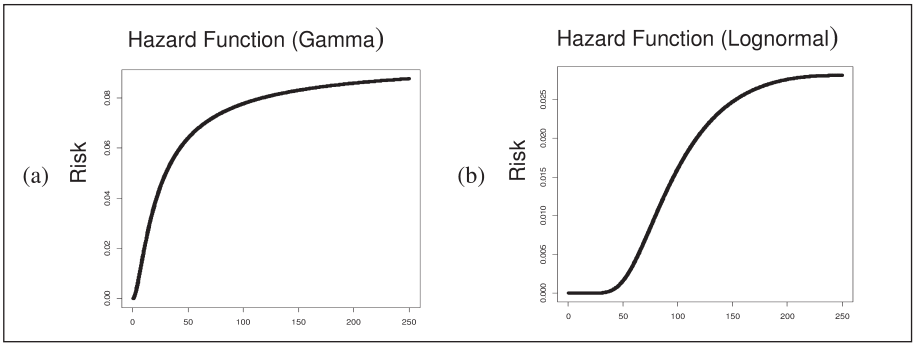


Figure 1. (a) Hazard function (Gamma PDF and survival function fitted only with simulated remission data); (b) Hazard function (Lognormal PDF and survival function fitted with simulated censored data).

FINAL CONSIDERATIONS

In the case of uncensored data, simulated survival function and Cox model are very similar, however in the case of censored data, the difference between the two models is clear and has impact in hazard functions which can be observed in Figure 1 (a) and (b).

Data analysis was performed using the R program language with *survival*, *fitdistrplus*, *Hmisc, GA* and *EQL* packages.

ACKNOWLEDGEMENTS

This work was partially sponsored by national founds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the project (PEst-OE/MAT/UI000 6/2014).

REFERENCES

Ardoino, I., Biganzoli, E. M., Bajdik, C., Lisboa, P.J., Boracchi, P., and Ambrogi, F. (2012). "Flexible Parametric Modelling of the Hazard Function in Breast Cancer Studies". *Journal of Applied Statistics*, 39 (7), 1409-1421.

Datta, D. (2013). "Uncertainty Modeling of Retention Function in Biokinetic Model Using Polynomial Chaos Theory-Development of Computational Algorithm". *International Journal of Mathematical* Archive (IJMA) ISSN, 2229-5046, 4 (4).

Dittus, R.S., Roberts, S.D., and Wilson, J.R. (1989). "Quantifying Uncertainty in Medical Decisions". *J.A.C.C.*, 14 (3), A23-A28.

Isukapalli, S.S., Roy, A., and Georgopoulos, P.G. (1998). "Stochastic Response Surface Methods (SRSMs) for Uncertainty Propagation: Application to Environmental and Biological Systems". *Risk Analysis*, 18 (3), 351-363.

Koerkamp, B.G., Weinstein, M.C., Stijnen, T., Heijenbrok-Kal, M.H., Hunink, M.M. (2010). "Uncertainty and Patient Heterogeneity in Medical Decision Models". *Medical Decision Making*, 30 (2), 194-205.

Leal, C., Oliveira T.A., and Oliveira A. (2014). "Stochastic Response Surface Methodology: A Study in the Human Health Area". In Simos, T.E., Psihoyios, G., Tsitouras, Ch., and Anastassi, Z. (eds.), *Numerical Analysis and Applied Mathematics ICNAAM 2014*, AIP Conference Proceedings, American Institute of Physics, to be published.

Oladyshkin, S.O., and Nowak, W. (2012). "Data-Driven Uncertainty Quantification Using the Arbitrary Polynomial Chaos Expansion". *Reliability Engineering & System Safety*, 106, 179-190.

Oliveira, T.A., Leal, C., and Oliveira, A. (2015). "Response Surface Methodology: A Review of Applications to Risk Assessment". In Kitsos, C., Oliveira, T., Rigas, A., and Gulati, S. (eds.), chapter XXIX, *Theory and Practice of Risk Assessment*, Springer Proceedings in Mathematics and Statistics (in press).

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2012.

Wisconsin Breast Cancer Prognosis Dataset. http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html#prog

Wolberg, W.H., Street W.N., and Mangasarian, O.L. (1999). "Importance of Nuclear Morphology in Breast Cancer Prognosis". *Clinical Cancer Research*, 5 (11), 3542-3548.

66. AN APPROACH FOR RISK ESTIMATION IN TERRITORIES/ PERILS UN-MODELLED BY THE INSURANCE INDUSTRY

D. Oramas Dorta
Guy Carpenter Analytics Dept. United Kingdom
Delioma.Oramas-Dorta@guycarp.com

ABSTRACT

Catastrophe risk models ("Cat models") are routinely used in the insurance industry to obtain quantitative measures of risk of occurrence of losses associated to perils such as earthquakes, hurricanes, etc. These measures are used by industry players as a basis for risk management. Building Cat models involves significant investment of time and resources; therefore they are commonly developed for territories/ perils where established insurance markets already exist. However, creation of new insurance markets in "non-modelled" territories/ perils needs, among other requisites, from some level of risk quantification to assess insurability and develop Cat insurance products. This work proposes a simple methodology to provide risk estimations for territories/perils where full-blown Cat models do not yet exist. It makes use of hazard data publicly available and a risk aggregation method based on Solvency II.

1. INTRODUCTION

Cat models consist of complex computer-assisted calculations with statistical, engineering and geographical bases, packaged into licensable software by specialised R&D companies. Outputs consist of probabilities of occurrence (expressed as Return Periods; R.P.) of different monetary loss amounts likely to be sustained by a portfolio of properties geographically spread across a territory, and potentially subject to the effects of the considered peril (e.g. earthquakes occurring in that territory).

This paper proposes a methodology for benchmark risk estimations in territories/perils where full-blown Cat models do not exist, to help assess insurability of existing Cat risks and the potential to develop Cat insurance markets. It uses hazard data publicly available and a risk aggregation methodology based on Solvency II Standard Formula (EIOPA, 2011).

2. DESCRIPTION OF METHODOLOGY

The methodology provides loss estimates for a Return Period of interest; for instance, the 1:200 year loss (the loss that is exceeded once in 200 years on average) is typically reported for earthquake risk in insurance. To explain the methodology, this paper focuses on estimation of the 1:200 yr. loss for a geographically-distributed portfolio of properties susceptible to seismic hazard. The estimation consists of the following two main steps:

1. **"Hot-spot"–level analysis:** estimation of risk for individual geographic concentrations of exposure ("hot spots") in a territory (e.g. 1:200 year loss for each key city in a seismically active territory). For cities x, y, z:

   - Lx = 1:200 yr. R.P. loss estimate for urban area/ hot spot "x"
   - Ly = 1:200 yr. R.P. loss estimate for urban area/ hot spot "y"
   - Lz = 1:200 yr. R.P. loss estimate for urban area/ hot spot "z"

   Estimation of the 1:200 yr. loss for each individual "hot spot" (L200_i) involves consideration of:

   a. The total monetary value of potentially exposed properties: "TIV_i".

   b. Hazard intensity level at the considered hot spot, for an equivalent R.P. (e.g. 1:200 yr. Peak Ground Acceleration for seismic hazard: PGA200_i), derived from publicly available hazard maps.

c. A damage function that provides a Loss Ratio (% of monetary value lost; LR_i) for properties exposed to a particular hazard level (e.g. to seismic intensity PGA200_i). The function chosen needs to be reasonably appropriate for the specific territory, peril and property type.

Therefore, to estimate the 1:200 R.P. loss for an individual "hot spot":

$$L_{200_i} = TIV_i \times LR_i$$

where LR_i = f (PGA200_i)

2. **Aggregated–level analysis:** estimation of aggregated risk for the overall portfolio (i.e. the set of "hot spots" considered jointly). Correlation among "hot spots" needs being considered, which mainly relates to geographical correlation in losses associated to events (such as earthquakes) impacting several "hot-spots" at the same time. A risk aggregation formula that makes use of a correlation matrix (CEIOPS, 2010) is used to estimate the aggregated 1: 200 yr. portfolio loss (LT):

$$L_T = \text{sqrt}\left(Lx^2 + Ly^2 + Lz^2 + 2*\rho x,y*Lx*Ly + 2*\rho x,z*Ly*Lz + 2*\rho y,z*Ly*Lz\right)$$

where:

L_x, L_y, L_z: individual, "hot spot" 1: 200 R.P. loss estimations ρ_x,y, $\rho_{x,z}$, $\rho_{y,z}$: loss correlation coefficients between pairs of "hot spots"

The Fifth Quantitative Impact Study for Solvency II (EIOPA, 2011) considers risk aggregation, providing double-entry matrices of correlation factors for different Cat perils in European countries. Each double-entry refers to a pair of geographic regions used for insurance aggregation purposes (Cresta zones) within each country (EIOPA, 2010). Since correlation factors depend on geographic distance between exposures (CEIOPS, 2010), this work has derived explicit relationships between distances (in Km.) among all pairs of Cresta zones within each country considered by QIS5, and correlations coefficients reported in the earthquake-specific aggregation matrices (EIOPA, 2010). Figure 1 (A) shows examples of such derived

relationships, which are used in the present methodology to read off correlation coefficients for any pair of "hot spots" given the distance between them; and consequently estimate LT by plugging those coefficients into the aggregation formula.

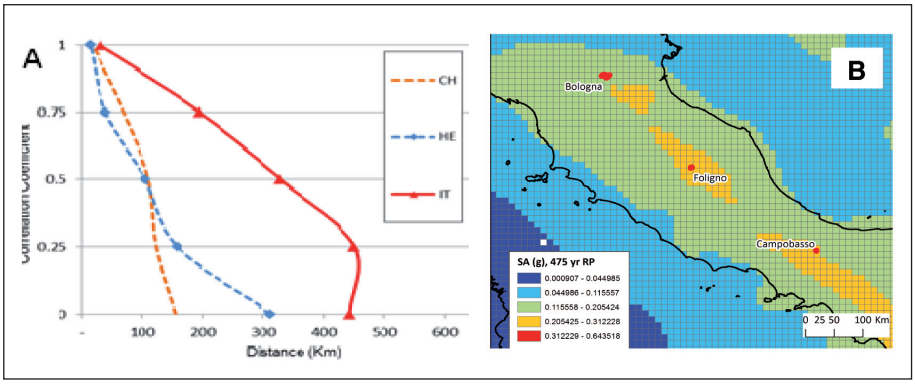


Figure 1. A: relationship between distance (in Km) and earthquake-related (average) correlation coefficients, derived from EIOPA (2010) Cresta-based correlation matrices for Switzerland (CH), Greece (HE) and Italy (IT). B: Location of test portfolio in Italy (red "hot spots") against EFEHR seismic hazard.

3. APPLICATION OF METHODOLOGY AND RESULTS

The methodology has been applied to estimate the 1:200 yr. R.P. loss of a portfolio of seismically-exposed residential buildings situated in Italy, a territory currently modelled by full-fledged Cat models. The aim is to compare results from the proposed methodology with risk estimates used in the insurance market. Both "hot-spot" and aggregated-level results are compared to the corresponding outputs from a commercial Cat model of Italy earthquake (from now on referred to as "CommCat").

Three different loss estimates are presented: one for methodology validation purposes (i.e. using hazard values and damage functions from "CommCat") and referred to as "Validation Estimation", and two other estimates: using independent seismic hazard data and damage functions from "CommCat" (Independent Estimation (1)), and independent seismic hazard data and a damage function from another modelled country (Independent Estimation (2)). "Independent Estimations"

represent more closely an actual application for un-modelled territories/ perils. The following describes the components of the test case for the "Validation" and "Independent" estimations (values summarized in Table 1):

a. A test portfolio composed of three "hot spots" (Italian cities of Bologna, Campobasso and Foligno) with concentrations of residential properties amounting to 10 Million aggregated currency units each (Figure 1 (B)).

b. The 1:200 yr. RP Spectral Acceleration ("SA"; seismic hazard intensity in g units) has been calculated for each "hot spot" from independent seismic hazard maps (SHARE project; Figure 1 (B)) (herein referred to as SA_indep). For validation, hazard values from "CommCat" have also been extracted (herein referred to as SA_Comm) for these "hot spots".

c. Damage functions providing a relationship between seismic hazard (SA in this case) at each city and damage ratios have been borrowed from existing commercial Cat models, as follows:

    - The damage function prescribed by "CatComm" for Residential buildings for Italy quake (i.e. Dam_It), for validation purposes.

    - A damage function from another modelled European country for Residential buildings (i.e. Dam_indep). This particular choice is for methodology demonstration purposes only; Dam_indep not necessarily being a close reflection of vulnerability of Italian Residential buildings. Application to un-modelled territories/ perils would entail research on independent damage functions representing the portfolio best.

The relationships between distance and correlation factors (Figure 1 (B)) have been used to estimate the aggregated 1:200 yr. RP portfolio loss, using the geographic distance among the three cities (Table 1). Two different relationships have been used: the Italy one for the Validation and Independent estimations; and the Greece one for Independent estimations (to explore the sensitivity of results to this choice).

	Validation Estimation	Independent Estimation (1)	Independent Estimation (2)
Property portfolio (Cb, Bol, Fol)	10 Mill. currency units/ city	10 Mill. currency units/ city	10 Mill. currency units/ city
SA200 (Cb, Bol, Fol)	SA_Comm: 0.110, 0.093, 0.160 (g)	SA_indep : 0.117, 0.103, 0.141 (g)	SA_indep : 0.117, 0.103, 0.141 (g)
LR = f(SA200) (Cb, Bol, Fol)	Dam_It:: 0.01764, 0.00589, 0.04394	Dam_It: 0.01764, 0.00589, 0.03042	Dam_indep: 0.02269, 0.01192, 0.03552
Geographic Distance (Cb-Bol, Cb-Fol, Bol-Fol)	487.2, 265.5, 225.3 (km)	487.2, 265.5, 225.3 (km)	487.2, 265.5, 225.3 (km)
Correlation coeff. (Cb-Bol, Cb-Fol, Bol-Fol)	Corr_It: 0.2500, 0.6174, 0.6926	Corr_It: 0.2500, 0.6174, 0.6926; Corr_He: 0, 0.14, 0.075	Corr_It: 0.2500, 0.6174, 0.6926; Corr_He: 0, 0.14, 0.075

Table 1. Components for 1:200 yr. Loss Estimations of test portfolio.

Table 2 provides "hot-spot" and "aggregated" results for "Validation" and "Independent" estimations, and their equivalent result from "CatComm".

Estimation Level	Validation Estimation	Independent Estimation (1)	Independent Estimation (2)	"CatComm" Result
1:200 yr. RP Loss ("hot spot"-level)	176,422 (Cb); 58,965 (Bol); 439,453 (Fol)	176,422 (Cb); 58,965 (Bol); 304,197 (Fol)	226,927 (Cb); 119,270 (Bol); 355,220 (Fol)	150,562 (Cb); 56,672 (Bol); 529,956 (Fol)
1:200 yr. RP Loss (aggregated level)	603,779 (Corr_It)	472,708 (Corr_It)/ 380,600 (Corr_He)	603,041(Corr_It)/ 469,928(Corr_He)	581,315/ 671,661 (agg.; Corr_It)

Table 2. Results of 1:200 yr. Loss Estimations of test portfolio.

Results for the "hot-spot" and aggregated levels are discussed, for each of the three Estimates/ sets of components considered (Tables 1 and 2):

1. "Hot spot" - level results

a. Validation estimation: comparison with "CatComm" results reveals differences ranging between 4% and 17%. These arise from the fact that it is not possible to retrieve the exact D.R. that "CatComm" applies), but approximated from analyses results.

b. Independent estimation (1): comparison with Validation results shows that using independent hazard values leaves loss results unaffected for Campobasso and Bologna. Losses for Foligno are 31% lower than the Validation estimate (SA_indep is lower than SA_Comm for this city).

c. Independent estimation (2): differences with the Validation estimation amount to 28.6%, 102% and (-) 19.2% for Campobasso, Bologna and Foligno, respectively. These are partly due to using independent hazard; but mainly, differences highlight sensitivity of results to choosing a damage function that is representative of the portfolio.

2. Aggregated 1:200 yr. loss results

a. Validation estimation: uses Validation results at "hot spot" level and correlation coefficients from the Italy relationship (Fig. 1 (B)). Results compare well with the "CatComm" Result (3.86% higher). Additionally, an equivalent estimation was made using "hot spot" level "CatComm" results, being 15.5% higher than the aggregated "CatComm" result straight from the model. The latter result is a more objective view of differences arising from the aggregation methodology.

b. Independent estimation (1): using Independent (1) "hot spot" level results and the Italy distance-correlation relationship, the aggregated estimate is 18.7% lower than the "CatComm" result. Another estimation using the Greece relationship results 34.5 % lower than "CatComm".

c. Independent estimation (2): using Independent (2) "hot spot" level results and the Italy distance-correlation relationship, the aggregated estimate is 3.74%

higher than "CatComm". Another estimation using the Greece relationship results 19.2 % lower than "CatComm".

4. CONCLUSIONS AND FURTHER WORK

Comparison between the "Validation Estimation" and "CatComm" results shows a good agreement at the "hot spot level" (differences between 4% and 17%, arising from extracting approximated D.R.s from "CatComm"). At the aggregated level, the difference of 3.9% comprehends the D.R.s approximations and the use of the aggregation methodology, which may compensate. To isolate the effects of the latter, we have used the "hot spot" level "CatComm" results and the Italy relationships, resulting in a Validation estimate 15.5% higher than the "CatComm" aggregated result. Independent Estimation (1) shows that using an alternative hazard source provides "hot spot" level results very similar to Validation results. The independent hazard source used must ideally be widely accepted by the scientific community. "Hot spot" results have been aggregated using distance-correlation relationships from Italy and from Greece, resulting in loss estimates 18.7% and 34.5% lower than the "CatComm" result, respectively. Therefore, choosing an appropriate (Italy-specific) distance-correlation relationship entails an improvement of 16% in this case. Independent Estimation (2) yields differences with respect to the Validation estimation between 102% and -19.2% at the "hot spot" level. This shows that the choice of damage function (borrowed from another modelled European country in this case) causes most sensitivity in results (see Independent Estimation (1)), and application of the methodology to a non-modelled territory entails careful selection of an independent damage function. The aggregated Independent Estimation (2) confirms that the choice of a representative distance-correlation relationship matters, but to a lesser extent than the choice of damage function.

The proposed methodology yields results comparable to those of full-fledged commercial Cat models, for the case study considered, but further validation using a wider range of case studies is needed. Extrapolating the technique to non-modelled territories requires using independent hazard data of sufficient quality, and careful selection of independent damage functions. Selection of a representative distance-correlation relationship causes sensitivity, but to a lesser extent than that of damage functions.

ACKNOWLEDGEMENTS

Special thanks to M. Weatherhead and G. Franco at Guy Carpenter.

REFERENCES

CEIOPS (2010). *Catastrophe Task Force Report on Standardised Scenarios for the Catastrophic Risk Module in the Standard Formula.* https://eiopa.europa.eu/CEIOPS-Archive/Documents/Reports/CEIOPS-DOC-79-10-CAT-TF-Report.pdf

EIOPA (2010). *QIS5: Parameters for Non-Life Catastrophe Risk*. https://eiopa.europa.eu/Publications/Quantitative%20impact%20studies/QI S5-CAT-risk-parameters_en.xls

EIOPA (2011). *Report on the fifth Quantitative Impact Study (QIS5) for Solvency II.* https://eiopa.europa.eu/Publications/Reports/QIS5_Report_Final.pdf

SHARE Project Data Portal. Accessible at: www.efehr.org

67. ALTERNATIVE METHODS OF ESTIMATING THE LONGEVITY RISK

A. Ornelas and C. Bolancé
Department of Econometrics, Riskcenter-IREA, University of Barcelona,
arelly.ornelas@ub.edu,
bolance@ub.edu

ABSTRACT

We focus on individuals over age 65 in order to quantify the longevity risk using the Value-at-Risk. We apply a robust estimation method for fitting parametric distributions. Alternatively, we also apply nonparametric and semi-parametric methods to evaluate the longevity risk. Finally we compare the longevity risk estimates with parametric, semiparametric and nonparametric methods.

1. INTRODUCTION

There is no certainty about the maximum age that an individual could attain; longevity, therefore, is considered a risk in some scenarios where increased longevity causes significant economic losses. For example, premiums in life insurance or annuities in pension plans are directly affected by the age of death.

We are interested in evaluating the risk of an individual surviving beyond the standard prevision. The difficulty in this case is that the information about the extreme age of death is scarce. The aim of this paper is to estimate the longevity risk and its trend depending on the given age of the individual.

In order to evaluate the longevity risk beyond age 65 (age of retirement), we will apply the concept of Value-at-Risk at α confidence level ($VaR\alpha,\ 0.95 \leq \alpha < 1$), to the

variable lifetime duration. VaR_{α} is commonly used to assess monetary losses. Longevity is not a monetary loss, but it is a risk, because it is directly related to activities that involve higher expenses when people get older. Precisely, the VaR_{α} reflects the number of years that only a small $(1-\alpha)\%$ fraction of the population survives. In Ornelas *et al.* (2013) longevity risk has been analysed under a parametric theory. These authors fitted a Weibull distribution to the lifetime duration variable and concluded that the fit is not entirely accurate.

The shape associated with lifetime duration variable is right skewed, i.e. deaths at ages near 65 years are frequent but extreme ages are reached quite infrequently. To find a distribution that fits the data in all their domain is not easy. Underestimation or overestimation is the main problem. Even when we select or simulate a large sample, the number of observed data at oldest ages is small. Then, we propose to use an alternative estimation method of the parametric distribution based on the class of minimum distance estimators, which has better robustness properties than maximum likelihood estimators (see Scott, 2001). Specifically, we obtain the parameters of the distribution by minimising an estimation of the distance between true cdf and estimated cdf.

Sample size is a key factor in determining the use of a nonparametric method versus a parametric distribution fit. A nonparametric alternative is the classical kernel estimator (CKE) that smooths the shape of the empirical distribution and "extrapolates" its behaviour when dealing with extremes, but when data are scarse CKE does not efficiently extrapolate the shape of the distribution above the maximum observed value in the sample. For this reason, we use a semiparametric method: the double transformed kernel estimator (DTKE) method. When using DTKE, we incorporate some bias in the estimation, but Alemany *et al.* (2013) proved that our proposed DTKE is a consistent estimator and its bias depends on the distance between the parametric cdf used in the transformation of the original data and the true cdf.

To analyze the trend of the longevity risk, we estimated the truncated cumulative distribution function, which allows us to obtain the risk conditional on the individual reaching a given age.

2. METHODOLOGY

Let X be a random variable that represents lifetime duration after age 65 and F (x) is the corresponding cdf. $F(x|X \geq a) = G(x)$ es the truncated cdf, where a $\geq$ 65. To estimate the risk of longevity we use a well-known measure of risk, the value-at-risk, that is equal to the α quantile of the distribution: $VaR\alpha(X) = G{-}1(\alpha) = inf\{x : G(x) \geq \alpha\}$, α $\in (0, 1)$.

2.1. Parametric distributions

To estimate the CDF, we propose using an extreme value distribution (EVD): the Lognormal and the Weibull distributions (that has a middle tail shape) and the modified Champernowne distribution (which is heavy-tailed), see, Buch-Larsen *et al.* (2005). To estimate parameters we can use the maximumlikelihood estimator (MLE., However, MLE has some problems: lack of robustness to outliers and model misspecification (see Scott, 2001). One solution consists of estimating the parameters minimising the distance between true cdf and estimated cdf. A criterion that is frequently used in the nonparametric methods involves minimising the mean integrated square error (MISE):

$$\text{MISE}\,[F_{\hat{\theta}}] = E\left[\int [F_{\hat{\theta}}(t) - F_{\theta}(t)]^2 \; dt\right], \tag{1}$$

where θ is a vector of parameters, F_{θ} is the true cdf and $F_{\hat{\theta}}$ is the estimated cdf. Altman and Leger (1995) proposed to estimate MISE from its discrete approximation and replacing the true cdf by the empirical cdf (Emp):

$\widehat{MISE}[F_{\hat{\theta}}] = n^{-1}\sum_{i=1}^{n}[F_{\hat{\theta}}(X_i) - F_n(X_i)]^2$, see also Bowman *et al.* (1998).

In order to compare MLE and $MISE$ criteria[1], in Figure 1 we plot the empirical cdf with parametric cdfs estimated by MLE and minimising $\widehat{MISE}$. In all figures, parametric estimates based on minimising $\widehat{MISE}$ are closer to the empirical cdf. For the Lognormal and the Champernowne distributions, the estimation based on minimising $\widehat{MISE}$ increases slower than the MLE estimation.

[1] For the graphic examples we use the data described in section 3.

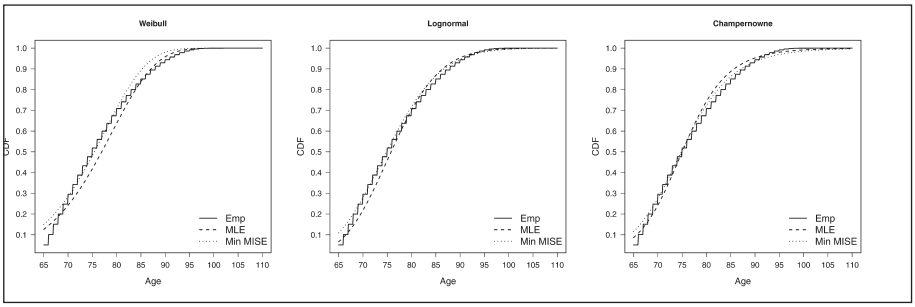


Figure 1. Lifetime variable distribution for male. Empirical cdf (solid line), MLE estimation (dashed line) and estimation based on minimizing $\widehat{MISE}$ (dotted line).

2.2. Nonparametric and semiparametric estimation

Nonparametric methods allow that the sample information itself draws the shape of the distribution. They provide an estimation for the value of the function at each point x in the domain of the random variable.

A natural nonparametric method to estimate cdf is the empirical distribution, but it is not efficient, see, Azzalini (1981) and cannot extrapolate beyond the maximum observed data point. This is especially troublesome when the data is scarce. In our data, the number of living individuals decreases when people get older and one may suspect that the probability of reaching older ages than the maximum observed age in the data sample is not zero. Furthermore, if $VaR\alpha$ is calculated empirically, an age in years is obtained. However, economic policies, pension and dependency regimes need to be planned considering all possible expenses; therefore, it is better to calculate the risk in units that are more precise than integer years.

An alternative to Emp is the classical kernel estimation (CKE) of $F(x)$, see Silverman (1986) and Alemany *et al.* (2013). In this paper we use the Epanechnikov kernel, $K(x) = 3/4(1 - x^2)$, *if* $|x| \leq 1$. The value of the smoothing parameter b considerably affects the results of the estimated cdf, especially when the goal is the estimation at a given point x for calculating the VaR. To calculate the smoothing parameter, we use the proposal of Alemany *et al.* (2013) that is based on the rule-of-thumb bandwidth proposed by Silverman (1986) and takes into account the existence of extreme

values. For estimating the VaR from CKE we need a numeric algorithm, for example the Newton-Raphson, that allows to solve the equation $\hat{G}(VaR_{\alpha}(X)) = \alpha$, where $\hat{G}$ is the CKE of the truncated cdf $\hat{G}$.

The CKE smooths the stepped shape of the empirical distribution; however, when data is rare the CKE does not allow us to accurately estimate the VaR. With the aim of improving the efficiency of CKE, Alemany *et al.* (2013) proposed the double transformed kernel estimator (DTKE). We can summarize the procedure in three steps:

- Step 1: Minimize MISE, by fitting a parametric distribution function, $F_{\hat{\theta}}$ and obtain the transformed data $U_1, \ldots, U_n$ that is close to a $Unif\,orm(0, 1)$.

- Step 2: Transform the uniform data using the inverse of the cdf of a *Beta (3,3),* $M(x) = \frac{3}{16}x^5 - \frac{5}{8}x^3 + \frac{15}{16}x + \frac{1}{2}$.

- Step 3: Obtain the CKE using the double transformed sample. The result is the called $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{M^{-1}\left(F_{\hat{\theta}}(x)\right) - M^{-1}\left(F_{\hat{\theta}}(X_i)\right)}{b}\right)$.

The DTKE is based on the fact that the cdf of a *Beta(3, 3)* can be estimated optimally using classical kernel estimation, see, Terrell (1990). An accurate optimal bandwidth for estimating *F(x)* is:

$$\widehat{b}_x = \left(\frac{m\left(M^{-1}\left(F_{\hat{\theta}}(x)\right)\right)\left(1 - \int_{-1}^{1} K^2(t)\,dt\right)}{4\left[\frac{1}{2}m'\left(M^{-1}\left(F_{\hat{\theta}}(x)\right)\right)\int_{-1}^{1} t^2 k(t)\,dt\right]^2}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{2}$$

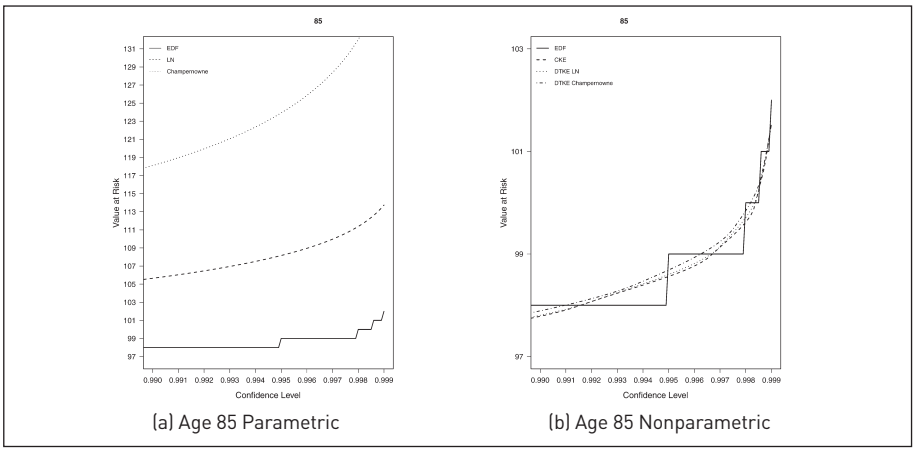
where $m = M'$ is the pdf of the *Beta (3, 3)* distribution. The DTKE improves the parametric fit by reducing its bias and improves the efficiency of the estimators that are only based on sample information, Alemany *et al.* (2013). We compare different DTKE obtained using different prior parametric distributions: Lognormal, Champernowne and Weibull. We can see that, when we use a heavier tailed initial parametric distribution, the estimated cdf increases more slowly than when using a lighter tail, i.e. we will estimate a larger longevity risk when using a heavy tail distribution than otherwise.

3. DATA ANALYSIS

The methodology described in section 2 has been used to estimate the longevity risk of the Mexican population covered by insurance. We use a simulated database from the mortality tables of those insured. The sources of information were the Mexican Association of Insurance Institutions (AMIS) and Mexican Insurers Association (AMA).

For each gender, we have simulated a data set with an initial population including N = 5000 individuals aged 65. In order to obtain the number of deaths, mortality rates were applied to the survivors at each age x. The sample size is equal to the sum of all the survivors at every age. We supposed that the survivors at 100 die randomly until the entire population disappears. Uniform distribution was used to decide whether a subject died or survived. The oldest ages were 107 for males and 112 for females.

We have estimated the parameters for all three distributions using MLE and minimum $\widehat{MISE}$. We have observed that for the Lognormal distribution and Weibull distribution the estimated parameters are similar for both estimation methods. The shape parameters of the Champernowne distributions estimated by minimum $\widehat{MISE}$ are lower than those estimated by MLE; consequently, the distributions have a heavy right tail. According to the minumun $\widehat{MISE}$, we conclude that the distribution that best fits the cdf of the lifetime variable is the Lognormal, followed closely by the Champernowne.



(a) Age 85 Parametric (b) Age 85 Nonparametric

Figure 2. Comparing different estimated VaR_α for males.

Figure 2 plot the curves associated with the estimated VaR for individuals aged 85, the results are obtained using parametric, semiparametric and nonparametric method. Parametric models provide a biased estimation of the VaR for all confidence levels, both models overestimate the longevity risk. When we use nonparametric and semiparametric methods the curves associated with estimated VaR are near the empirical VaR and they smooth the empirical VaR, although there exist some differences between the three plotted curves. In general, the largest risk is obtained with DTKE with the Champernowne distribution as initial transformation.

4. CONCLUSION

This paper has shown how the variable lifetime duration can be modeled following different approaches: parametric and nonparametric statistics. We used the $\widehat{MISE}$ to compare several cumulative distribution function estimates regardless of their method of estimation. Additionally, $\widehat{MISE}$ was used in the parametric framework as method of estimation. $\widehat{MISE}$ converges quickly to a global minimum when maximun likelihood parameters are used as starting values. The advantage of using $\widehat{MISE}$ as an estimation method is to add flexibility to the fitted cdf. Even though the Lognormal distribution is the best fit for males, this estimation does not recreate the real behavior of the observed data in the tails. When the confidence level increases, the VaR_α goes to ages that are unlikely to be reached, then, VaR_α does not reflect the natural age limit. The semiparametric estimation of the longevity risk that we propose in this paper is an excellent alternative, on the one hand it improves the efficiency of nonparametric estimation and, on the other hand, it addresses the possible large bias in finite sample and the possible inconsistency of parametric estimation.

ACKNOWLEDGMENTS

We acknowledge the support received from Spanish Ministry and FEDER grant ECO2013-48326-C2-1-P. We also thank members and affiliates of Riskcenter at the University of Barcelona for their comments and suggestions.

REFERENCES

Alemany, R., Bolancé, C., and Guillén, M. (2013). "A Nonparametric Approach to Calculating Value-at-Risk". *Insurance: Mathematics and Economics*, 52, 255-262.

Altman, N., and Leger, C. (1995). "Bandwidth Selection for Kernel Distribution Function Estimation". *Journal of Statistical Planning and Inference*, 46, 195-214.

Azzalini, A. (1981). "A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method". *Biometrika*, 68, 326-328.

Bowman, A., Hall, P., and Prvan, T. (1998). "Bandwidth Selection for the Smoothing of Distribution Functions". *Biometrika*, 85, 799-808.

Buch-Larsen, T., Nielsen, J.P., Guillén, M., and Bolancé, C. (2005). "Kernel Density Estimation for Heavy-Tailed Distributions Using the Champernowne Transformation". *Statistics*, 39, 503-516.

Ornelas, A., Guillén, M., and Alcañiz, M. (2013). "Implications of Unisex Assumptions in the Analysis of Longevity for Insurance Portfolios". *Lecture Notes in Business Information Processing*, 145, 99-107.

Scott, D.W. (2001). "Parametric Statistical Modeling by Minimum Integrated Square Error". *Technometrics*, 43, 274-285.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Terrell, G.R. (1990). "The Maximal Smoothing Principle in Density Estimation". *Journal of the American Statistical Association*, 85, 470-477.

68. EXTREME VALUE THEORY IN COMPUTER SCIENCES: THE CASE OF EMBEDDED SAFETY-CRITICAL SYSTEMS

J. Abella
Barcelona Supercomputing Center (BSC), Spain,
jaume.abella@bsc.es

J. del Castillo and M. Padilla
Universitat Autònoma de Barcelona, Departament de Matemàtiques, Spain, castillo@mat.uab.cat, mpadilla@mat.uab.cat

F.J. Cazorla
Barcelona Supercomputing Center (BSC) and Spanish National Research Council (IIIA-CSIC), Spain,
francisco.cazorla@bsc.es

ABSTRACT

A new set of techniques have been proposed to predict the worst-case execution time (WCET) behavior of programs. Those techniques use extreme value theory (EVT) as their main prediction method. In this paper the EVT is used to estimate a high quantile, that is the value at risk (VaR), for several distributions observed for a set of representative programs.

1. INTRODUCTION

The computing systems deployed in planes, cars, satellites or trains have strong requirements on guaranteeing that the longest execution time a program can take is bounded. These computing systems are called embedded real-time safety-critical systems. A number of methods to estimate the WCET exists, each

one with its own set of constraints, Wilhelm *et al.* (2008). Recently, measurement-based probabilistic timing analysis (MBPTA) has emerged as a powerful approach to derive those WCET estimates in powerful internally-complex computing systems, Cucu-Grosjean *et al.* (2012). MBPTA relies on platforms whose execution time can be modelled with a random variable (r.v.) and thus, by collecting execution time measurements conveniently, Cazorla *et al.* (2013), a correct application of EVT allows estimating the distribution of the highest execution times. In this paper we analyze some of the assumptions made in such application and their implications using some of the latest findings in the field of EVT, Castillo *et al.* (2014) and Castillo and Serra (2014). In particular, we focus on the analysis of single-path automotive applications through the use of the EEMBC Autobench suite, Poovey (2007). We consider a hardware architecture implementing some high-performance features such as multicores and cache hierarchies analogous to that in Slijepcevic *et al.* (2014), which has been modelled by means of a cycle-accurate simulator based on SoCLib, see SoClib.

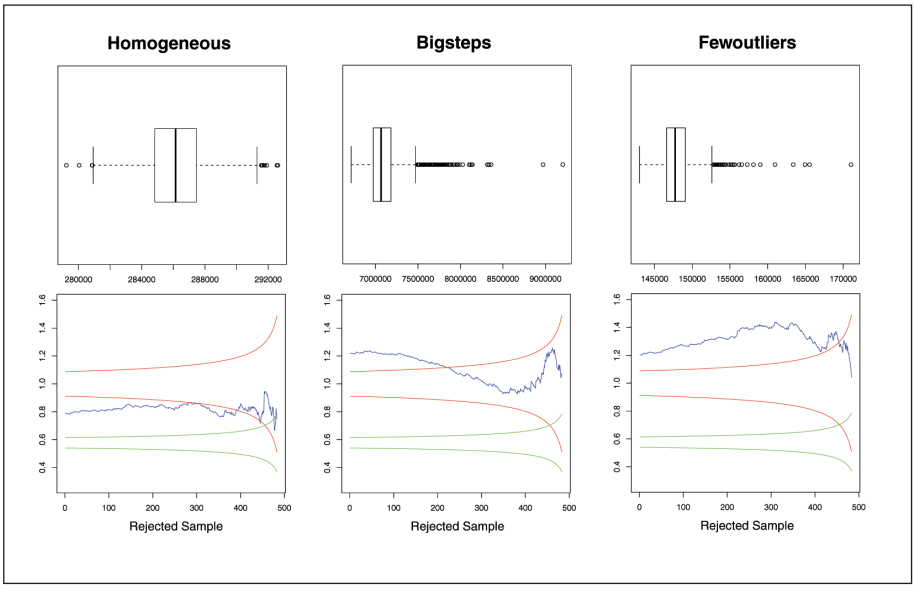


Figure 1: Box-plot and CV-plot for the last 500 order statistics of the samples. The red and green lines correspond to the 95% confidence interval for exponential and uniform distributions, respectively.

This work analyzes 16 data sets, each with 1,000 observations of the execution time, in cycles, for different programs. The box-and-whisker plot is used, where $IQR = (Q_3 - Q_1)$ is the interquartile range, W_1 is the lowest datum still within $1.5IQR$ of the Q_1, and W_2 the highest datum still within $1.5IQR$ of the Q_3 (often called the Tukey boxplot). The 16 datasets show significantly more values beyond the interval (W_1, W_2), these values exceed almost all W_2, proportions ranging from 0.8% to 8.7% of the sample and the distance between W_2 and the maximum ranging from 0.5 to 32.2 in terms of IQR. From now on, for space reasons, only the results of three representative sets are presented, called: Homogeneous, Bigsteps and Fewoutliers.

2. GEV AND GPD

The EVT has two main approaches: block maxima models and threshold exceedance models. The more classical parametrical approaches are generalized extreme value (GEV) for maxima and generalized Pareto distribution (GPD) for exceedances. In both cases the asymptotic results are derived under the assumption of independence, although that can be applied with little change in the case of stationary processes (Coles, 2001, Ch 5). GEV only needs the independence of the maximums that can occur in stationary cases, where the observations can be time-dependent, but modeling only block maxima is a wasteful approach to extreme value analysis if other data on extremes are available and independence can be assumed. A popular and simple test of independence uses the first autocorrelation coefficient rejecting at the 5% significance level if $\sqrt{n}\,|\hat{\rho}_1| > 1.96$.

	Homogeneous		Bigsteps		Fewoutliers	
h	LB	p -value	LB	p -value	LB	p -value
10	10.72	0.380	8.69	0.562	9.95	0.444
20	16.15	0.708	14.32	0.814	26.76	0.142

Table 1. Results of Box-Ljung test for the representative samples.

The test applies the result $\sqrt{n}\varrho_k \sim N(0,1)$ for the autocorrelation coefficient at lag k, approximately, when H_0 is true. When many coefficients are considered some will

probablily be significant even if H_0 is true. A more natural way to combine h coefficients into a single test statistic is the Ljung–Box test given by

$$LB = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where n is the sample size, and $\hat{\rho}_k$ is the sample autocorrelation at lag k. Under the hypotesis of independence, on the assumption that the data at least has finite fourth moment, the statistic follows a χ^2_h distribution. The LB statistic is frequently used in general time series research to test zero autocorrelation. Others tests, like the runs test, essentially only consider the lag 1 autocorrelation, its disavantatge is low test power caused by losing information. Table 1 shows the test results applied to the representative data, with 10 and 20 lags. In all cases the p-value is not significant, so there is no evidence against that data are independent. This leads us to use threshold exceedance with GPD for the right tail.

3. PEAK OVER THRESHOLD (PoT) METHOD

There are two important points to consider in PoT: the threshold selection and the parameter estimation, see Coles (2001). The threshold for the excesses is usually chosen by a graphical method, this paper proposes to use the CV-plot, which was introduced by Castillo *et al.* (2014).

Let X be a continuous non-negative r.v. with distribution function $F(x)$. For any threshold, $t > 0$, the r.v. of the conditional distribution of threshold excesses $X - t$ given $X > t$, denoted $X_t = (X - t \mid X > t)$, is called the residual distribution of X over t. The cumulative distribution function of X_t, $F_t(x)$, is given by

$$1 - F_t(x) = (1 - F(x+t)) / (1 - F(t))$$

The quantity $M(t) = E(Xt)$ is called the residual mean and $V(t) = var(X_t)$ the residual variance. The residual coefficient of variation is given by

$$\mathrm{CV}(t) \equiv \mathrm{CV}(X_t) = \sqrt{V(t)}/M(t);$$

like the usual CV, the function $CV(t)$ is independent of scale, and the function $t \longrightarrow CV(t)$ is constant if and only if the distribution of X is GPD.

Sample size	m	$T_m/(m+1)$	p-value	Exponential	$\hat{\xi}$
300	5	3.38	0.016	No	-0.24
200	4	1.94	0.068	No	-0.23
150	4	2.45	0.032	No	-0.31
100	3	1.28	0.164	Yes	-0.26
50	2	1.03	0.217	Yes	-0.33

Table 2. Testing exponentiality by Tm and MLE for the shape parameter for Homogeneous.

The CV-plot has some advantages over ME-plot: first, it does not depend on the scale parameter; second, detecting constant functions is easier than linear functions, since linear functions are defined by two parameters and the constants by only one. The uncertainty is essentially reduced from three to one single parameter.

While the basic model for probability distributions is the normal distribution, for the threshold exceedance is the exponential distribution, in this case the shape parameter for the GPD is $\xi = 0$. It will be interesting, therefore, to contrast if this model is the most appropriate to fit the data. In this sense, the methods to contrast exponentiality versus polinomial tails developed in Castillo *et al.* (2014) will be used.

The contrast statistic is

$$T_m = n\sum_{k=0}^{m} 2^{-k}\,(cv(q_k) - 1)^2$$

where q_k are the quantiles corresponding to the last $n/2^k$ elements.

The exponentiality of the data has been contrasted, using different thresholds in the same way as Ljung-Box test uses several autocorrelations, with an appropriate value of m to study the right tail, and the ξ parameter has been estimated by MLE. The p-values have been calculated by simulation of 104 samples with the corresponding size in each case.

Table 2 shows exponentiality from the last 100 observations for Homogeneous, but the MLE estimation is far from the value $\xi = 0$. The CV-plot will be used to decide between these two options.

Figure 1 shows the box-plot and the CV-plot for the representative samples. It can be observed, in the CV-plot of Homogeneous sample (on the left), how the graph is within the exponentiality interval at the end, but there seems to be a constant behavior with CV less than 1 from the latest 300 observations indicating a distribution with short tail with parameter ξ near -0.25.

In Bigsteps, exponentiality is accepted from the 200 latest data, and the MLE estimator is close to 0, so the exponential model seems appropriate. Exponentiality is not accepted in any case to Fewoutliers, and the MLE estimator is greater than 0, so the exponential model does not seem appropriate. The CV-plot in Figure 1, on the right, indicates a distribution with heavy tails with ξ near 0.23 for the last 100 observations.

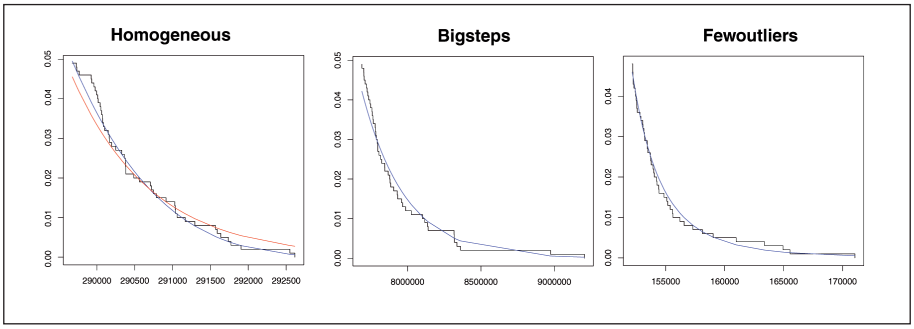


Figure 2. Empirical CCDF from quantile 95 for representative data. The blue lines correspond to the CCDF with the estimated parameters and the red line corresponds to the exponential approximation for Homogeneous.

	u	$\hat{\xi}$	VaR 95%	VaR 99%	VaR 99.9%
Hom.	287842	-0.25	289664.1 (50)	291121.3 (10)	292408.8 (2)
Big.	7228374	0	7638979 (57)	8115678 (10)	8797678 (2)
Few.	150462	0.23	151970.5 (54)	156556.2 (9)	166905.8 (1)

Table 3. Different VaRs for the thresholds and estimated parameters (in brackets the number of sample values above).

Figure 2 shows the fit of the proposed distribution to the data after estimate the scale parameter ψ by MLE. The fits seem good to use them to predict. Finally, some high quantiles for each sample are presented. In terms of risk, this concept is expressed as the value at risk (VaR). For a small p, $VaR_p = v$ if and only if $1 - F(v) = p$.

Table 3 shows the differents values for each proposed threshold and estimated parameter. The values in brackets, which exceed the VaR, are reasonable with the percentage of the values that should be above it, 50 for 95%, 10 for 99% and 1 for 99.9% for 1,000 observations datasets.

4. CONCLUDING REMARKS

1. Using a powerful non-correlation test under several scales, the datasets considered look like an independent sequence, hence GPD is recommended.

2. Several types of distributions arise in the examples, from light tailed to very heavy tailed distributions.

3. The exponential distribution is the equilibrium point in the study of the tails and has to be considered as the null hypothesis testing against polynomial or normal tails.

4. The calculation of VaR provides similar results adjusting the GPD distribution and exponential distribution when the null hypothesis is accepted. This shows the robustness of the procedure.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under the PROXIMA Project (grant agreement 611085). This work has also been partially supported by the Spanish Ministry of Science and Innovation under grants MTM2012-31118 (2013-2015) and TIN2012-34557. Jaume Abella is partially supported by the Ministry of Economy and Competitiveness under Ramon y Cajal postdoctoral fellowship number RYC-2013-14717.

REFERENCES

Castillo, J.D., Daoudi, J., and Lockhart, R. (2014). "Methods to Distinguish Between Polynomial and Exponential Tails". *Scandinavian Journal of Statistics*, 41, 382-393.

Castillo, J. D., and Serra, I. (2014). "Likelihood Inference for Generalized Pareto Distribution". *Computational Statistics and Data Analysis*. DOI: http://dx.doi.org/10.1016/j.csda.2014.10.014.

Cazorla, F.J., Vardanega, T., Quinones, E., and Abella, J. (2013). "Upperbounding Program Execution Time with Extreme Value Theory". *Worst-Case Execution Time Workshop*.

Coles, S. (2001). *An Introduction to Statistical of Extremes Values*. Springer, London.

Cucu-Grosjean, L., Santinelli, L., Houston, M., Lo, C., Vardanega, T., Kosmidis, L., Abella, J., Mezzetti, E., Quinones, E., and Cazorla, F.J. (2012). "Measurement-Based Probabilistic Timing Analysis for Multipath Programs". *Proceedings of the 2012 24th Euromicro Conference on Real-Time Systems*, ECRTS '12, 91-101.

Poovey, J. (2007). *Characterization of the EEMBC Benchmark Suite*, North Carolina State University.

Slijepcevic, M., Kosmidis, L., Abella, J., Quiñones, E., and Cazorla, F.J. (2014). "Time-Analysable Non-Partitioned Shared Caches for Real-Time Multicore Systems". *Proceedings of the 51st Annual Design Automation Conference*, DAC 14, 1-6.

SoCLib. http://www.soclib.fr/trac/dev.

Wilhelm, R., Engblom, J., Ermedahl, A., Holsti, N., Thesing, S., Whalley, D., Bernat, G., Ferdinand, C., Heckmann, R., Mitra, T., Mueller, F., Puaut, I., Puschner, P., Staschulat, J., and Stenström, P. (2008). "The Worst-case Execution-time Problem – Overview of Methods and Survey of Tools". *ACM Trans. Embed. Comput. Syst.*, 7(3), 36-53.

69. CONSIDERING RISK ISSUES IN THE UNCAPACITATED FACILITY LOCATION PROBLEM WITH STOCHASTIC COMPONENTS

A. Pagès-Bernaus and Á.A. Juan
Computer Science Department, IN3 - Open University of Catalonia, Barcelona, Spain, {apagesb@uoc.edu, ajuanp@uoc.edu}

H. Ramalhinho
Department of Economics and Business, Univ. Pompeu Fabra, Barcelona, Spain, helena.ramalhinho@upf.edu

ABSTRACT

The uncapacitated facility location problem is a classical problem where an undetermined subset of facilities are chosen to serve a set of customers such that some goal measure is optimized. The goal measure usually is the minimization of the sum of the (annualized) fixed setup costs plus the variable costs of serving the customers. In this paper we extend this problem to the case where the costs of serving the customers are random variables. A two-stage stochastic programming model is put forward to analyze the risk involved in considering using a simpler model with only average values.

The model is tested on several small benchmark cases from the Bilder-Krarup test set and solved optimally with Ilog-Cplex.

1. INTRODUCTION

The uncapacitated facility location problem (UFLP) is a classical problem presented as a plant location problem back in the 1960's. Given a set of potential locations and a set of customers to be served, the UFLP determines the optimal set of facilities (locations) such that all customers are served meeting the minimum overall cost, *i.e.* the sum of fixed (usually sunk) costs plus the sum of the costs of serving the facilities. In the uncapacitated version of the model it is assumed that the facilities can serve all the demand that the customers require and that this demand do not impact the serving cost or that the demand is known and therefore the incurred cost.

With the aim of incorporating more realistic components into this formulation, this paper extends the original formulation to the situation where the serving costs are stochastic, either because the demand is a random variable or because the travelling times can have significant variability. We still assume the property that the facilities can serve all the demand.

When uncertainty is taken into account, one needs to think about the timing between the uncertainty resolves and the set of decisions that can be taken before and after such information is revealed. In the UFLP with stochastic costs, we have two types of decisions. One is to decide which facilities are selected and the other is to assign customers to the cheapest facility. This situation is well represented by a two-stage stochastic programming model, where uncertainty appears only in the objective function. The decision to open a facility happens in the first stage before knowing the actual service cost. In the second stage, once the service costs are known, the assignment of customers to open facilities is done. In this context we assume that the selection of the facilities to open is a here-and-now decision that has to be taken in the first stage, while the assignment of the customers to the open facilities can be done after knowledge about the uncertain parameter is revealed, being the recourse action.

The interest on solving stochastic facility location problems dates back to 1976 when Balachandran and Jain developed several methodologies for solving the facility location problem when customers demand are stochastic (Balachandran and Jain,

1976). At the same time and for the same problem, Jucker and Carlson presented a mean-variance formulation to incorporate the risk-averse nature of a firm (Jucker and Carlson, 1976). Later other works presented several models and fields of application (see Snyder, 2004, for a review of models). More recently, Schutz *et al.* (2008) use a stochastic programming model for solving a facility location problem where fixed costs are non-liner and stochasticity appears both in costs and demand.

2. MATHEMATICAL MODEL: FROM DETERMINISTIC TO STOCHASTIC

The deterministic model of the UFLP can be formulated as follows:

$$\min_{x,y} Z_{det} = \sum_{i\in\mathbb{F}} f_i y_i + \sum_{i\in\mathbb{F}}\sum_{j\in\mathbb{C}} c_{ij}x_{ij} \tag{1a}$$

$$\text{s.t.} \sum_{i\in\mathbb{F}} x_{ij} = 1 \qquad ,\forall j \in \mathbb{C} \tag{1b}$$

$$x_{ij} \leq y_i \qquad ,\forall i \in \mathbb{F}, j \in \mathbb{C} \tag{1c}$$

$$0 \leq x_{ij} \leq 1,\ y_i \in \{0,1\} \quad ,i \in \mathbb{F}, j \in \mathbb{C}; \tag{1d}$$

where $\mathbb{F}$ is the set of facilities, $\mathbb{C}$ is the set of customers to be served and $f_i \geq 0$ is the (annualized) fixed cost of opening a facility at location $i \in \mathbb{F}$. The cost of supplying all demand of customer j from facility i is computed as the unitary transportation cost $t_{ij} \geq 0$ times the demand of customer j (in units of product) $d_j \geq 0$, here denoted as $c_{ij} = t_{ij}\, d_j$.

The decision variables are the binary variable y_i that indicate whether a facility i is opened or not, and x_{ij} accounts for the fraction of demand of customer j satisfied by facility i. Since we are in an uncapacitated version of the problem, this variable at the solution will be either 0 or 1, given that all the demand of a particular customer will be served by the facility which has the lowest cost c_{ij}.

The overall objective (1a) is to minimize the combination of fixed and variable costs. Here we assume that no economies of scale happen and the transportation costs behave linearly. Constraints (1b) ensure that the whole demand is served to each customer and constraints (1c) guarantee that customers are served only from open facilities.

It is common in practice that some of the data is a random variable whose distribution is known or can be estimated. Here we assume that the serving costs are stochastic, as it happens when goods are transported and the serving costs are directly related to the travelling times. In this situation, the capacity of the facilities and therefore the fixed costs are not affected. Assuming that the main decision to be taken is where to locate the facilities and that the assignment of the customers to the selected locations can be done after the knowledge about the travelling costs is known, a two-stage stochastic programming model perfectly fits. The randomness of the travelling costs are represented by means of a finite set of scenarios.

In a context with random variables it is crucial to consider the effect of variability, which leads to the concept of risk. One of the main approaches is a mean-risk model, where both the mean and a variability-related measure appears in the objective function. A popular asymmetric risk measure is the Conditional Value-at-Risk (CVaR) (see for example, Ogryczak and Ruszczynski, 2002).

The CVaR at level α is defined as follows (Rockafellar and Uryasev, 2000):

$$\text{CVaR}_{\alpha}(W) = E(W|W \geq VaR_{\alpha}(W)) \tag{2}$$

The CVaR minimizes the expected value of the variable costs *(W)* in the $(1 - \alpha)\%$ worst cases, where $\alpha \in (0, 1)$ is a preselected probability.

A mean-risk objective function allows to express the decision maker attitude towards risk by fixing a λ parameter to weight in different manner the overall expected cost versus the risk measure. Moreover, the user needs to specify the α-percentile of good solutions that is willing to accept. The higher α is the more solution accepts as good (given that the aim here is to minimize the overall total costs).

The extension to a two-stage stochastic programming model with a mean-risk model is presented in (3). Given that the decision of which facilities serve to which customers depends on the scenario, these variables are extended with a new supraindex $x_{ij}^{k}, k \in \mathbb{S},$, where $\mathbb{S}$ is the set of scenarios. Each scenario has a probability π^{k} and $\sum_{k \in \mathbb{S}} \pi^{k} = 1$. The first term of the objective function (3a) computes the average objective function value while the second term expresses the CVaR. For

computing the CVaR (Schultz and Tiedemann, 2006) a set of new variables is used, π^k, $k \in \mathbb{S}$, which accounts for the amount by which each scenario exceeds the Value-at-Risk, if so. The Value-at-Risk is the free variable η.

$$\min_{y,x,v,\eta} Z_{stoch} = (1-\lambda)\left\{\sum_{i\in\mathbb{F}} f_i y_i + \sum_{k\in\mathbb{S}} \pi^k \sum_{i\in\mathbb{F}}\sum_{j\in\mathbb{C}} c_{ij}^k x_{ij}^k\right\} + \lambda\left\{\eta + \frac{1}{1-\alpha}\sum_{k\in\mathbb{S}} \pi^k v^k\right\} \tag{3a}$$

$$\text{s.t.} \sum_{i\in\mathbb{F}} x_{ij}^k = 1 \qquad , j \in \mathbb{C}, k \in \mathbb{S} \tag{3b}$$

$$x_{ij}^k \le y_i \qquad , i \in \mathbb{F}, j \in \mathbb{C}, k \in \mathbb{S} \tag{3c}$$

$$v^k \ge \sum_{i\in\mathbb{F}}\sum_{j\in\mathbb{C}} c_{ij}^k x_{ij}^k - \eta \qquad , k \in \mathbb{S} \tag{3d}$$

$$x_{ij}^k \ge 0,\ y_i \in \{0,1\} \qquad , i \in \mathbb{F}, j \in \mathbb{C}, k \in \mathbb{S} \tag{3e}$$

$$\eta \in \mathbb{R}, v^k \ge 0 \qquad , k \in \mathbb{S}, \tag{3f}$$

3. RESULTS

Model (3) has been solved for the test case B1.1 from the Bilder-Krarup benchmark (Bilde and Krarup, 1977) which has 50 facilities and 100 customers. A lognormal distribution has been considered as the distribution of the costs. The scenario tree has been generated with a Monte Carlo simulation. Although this is not the method which provides the smallest tree, it asymptotically converges to the original distribution. After calibrating the adequate size of the tree, it has been found that 200 scenarios provides stability of the objective function value.

Each scenario consists of a matrix $|\mathbb{F}| \times |\mathbb{S}|$, where each element of the matrix c_{ij}^k is sampled from a lognormal distribution, X, with expected value $E(X) = c_{ij}$ (the original serving cost value) and the variance is proportional to the average, $V(X) = \kappa E(X)$. The original costs in this example c_{ij} vary between 0 and 999, and the values of the samples vary between 0 and 1180.

The test cases have been implemented in Java and solved with Cplex 12.1 on a laptop with 4 cores at 2.6GHz.

Table 1 shows the results of solving case B1-1 of the Bilder-Krarup benchmark for the deterministic case and for the stochastic case generated with $\kappa = 2$. The first column shows the stochastic objective function value and the second shows the number of scenarios used. It follows the α and λ parameters. In the following columns we compute the fixed costs and the expected serving costs, which added together gives the total costs. Finally, we show the number of facilities open in each solution. The solution time ranges from 1 second for the deterministic case to around three hours for the cases with $\lambda = \{0.5, 0.75\}$.

As expected, the more weight we put on the λ parameter (meaning that we are more concerned on avoiding solutions with high service costs), the more facilities are open. This bring an increase of the fixed costs but it reduces the expected cost of the variable costs. Therefore, avoiding risk is hedged by increasing the number of open facilities. This type of solution is not found if the deterministic solution with the mean-risk objective function is solved. Given a set of open facilities (found by the solution of the optimization model) one can simulate a large number of scenarios (for example 10,000) and assess more accurately the quality of the solution in terms of variable costs. Figure 1 shows the histogram of such runs for the deterministic case, the stochastic case with $\lambda = 0$ (i.e., minimizing only the expected value), and for $\lambda = \{0.25, 0.5, 0.75\}$. The risk-neutral solution ($\lambda = 0$) shows the same result as the deterministic case. This fact makes sense because given that the expected values are the same for both cases (deterministic and stochastic with $\lambda = 0$), the set of open facilities is the same (although the objective value has small variations). The histograms with some level of risk aversion show clearly how the variable/travelling costs diminishes, although the overall total cost increases (higher fixed costs are to be added, see Table 1).

Z^*_{stoch}	$\|\mathbb{S}\|$	α	λ	Fixed costs	Variable costs	Total costs	Num. Open Facilities
23468,0	1	-	0	0 9779	13689	23468,0	5
23400,3	200	0,95	0	0 9779	13621	23400,3	5
21003,2	200	0,95	0,25	11645	12183	23828,4	6
17952,8	200	0,95	0,5	14606	10502	25108,2	7
13669,7	200	0,95	0,75	18861	8738	27599,5	9

Table 1. Results from solving the case B1.1 of the Bilder-Krarup benchmark for different values of λ and $\kappa = 2$.

4. CONCLUSION

The uncapacitated facility location problem is a classical NP-hard problem which can be solved with off-the-shelf solvers for small to medium sized problems. A variant of this model is presented where the variable costs of serving the clients from the open facilities is considered as a random variable. This variation, yet minimal, allows us to clearly interpret the effects of having variability in such point and assess the risk of taking a decision. In order to account for the risk aversion of the decision taker, a mean-risk function in a two-stage stochastic programming model is put forward, and an analysis of the variation of such parameter is shown.

The analysis is done on an instance of the Bilder-Krarup benchmark and will be extended to other cases. Stochasticity on other parameters of the model (such as the fixed costs) and other risk measures are underway.

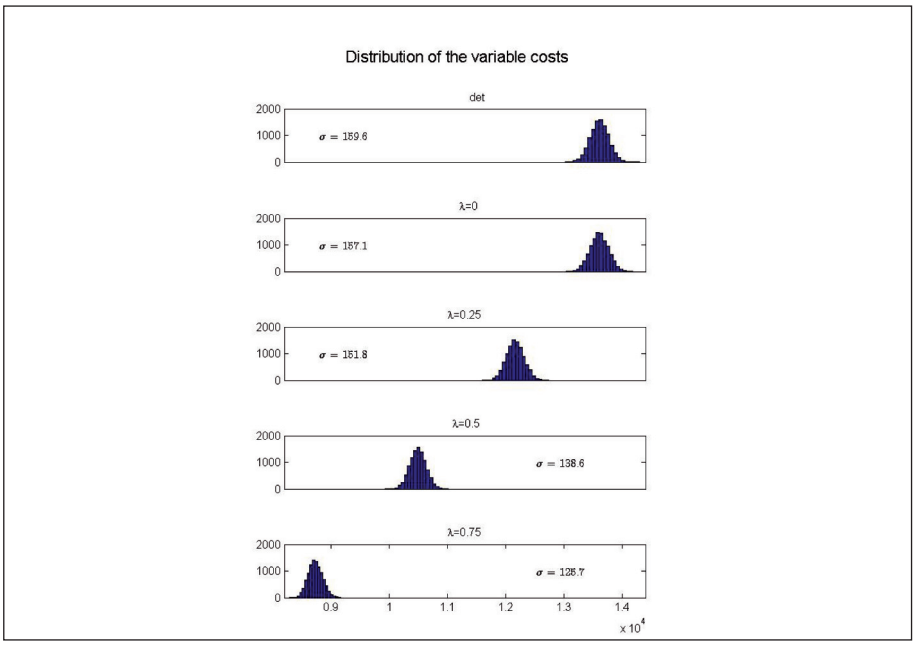


Figure 1. Histogram of the variable costs for case B1.1 (κ = 2) varying the λ parameter.

ACKNOWLEDGMENTS

This research has been partially supported by the Spanish Ministry of Economy and Competitiveness (grant TRA2013-48180-C3-3-P) and FEDER.

REFERENCES

Balachandran, V., and Jain, S. (1976). "Optimal facility location under Random Demand with General Cost Structure". *Naval Research Logistics*, 23, 421-436.

Bilde, O., and Krarup, J. (1977). "Sharp Lower Bounds and Efficient Algorithms for the Simple Plant Location Problem". *Annals of Discrete Mathematics*, 1, 79-97. Benchmark downloaded from http://resources.mpi-inf.mpg.de/departments/d1/projects/benchmarks/UflLib/packages.html (Jaunary 2015).

Jucker, J.V., and Carlson, R.C. (1976). "The Simple Plant-Location Problem under Uncertainty". *Operations Research*, 24, 1045-1055.

Ogryczak, W., and Ruszczynski, A. (2002). "Dual Stochastic Dominance and Related Mean-Risk Models". *SIAM Journal on Optimization*, 13 (1), 60-78.

Rockafellar, R.T., and Uryasev, S. (2000). "Optimization of Conditional Value-at-Risk". *Journal of Risk* , 2, 21-41.

Schultz, R., and Tiedemann, S. (2006). "Conditional Value-at-Risk in Stochastic Programs with Mixed-Integer Recourse". *Mathematical Programming*, 105 (2-3), 365-386.

Schutz, P., Stougie, L., and Tomasgard, A. (2008). "Stochastic Facility Location with General Long-Run Costs and Convex Short-Run Costs". *Computers and Operations Research*, 35 (9), 2988-3000.

Snyder, L. V. (2004). "Facility Location under Uncertainty: A Review". *IIE Transactions*, 38, 547-564.

70. ON SEVERAL GENERALIZATIONS OF CLASSICAL TAIL INDEX ESTIMATORS

V. Paulauskas
Vilnius University, Department of Mathematics and Informatics, Lithuania,
vygantas.paulauskas@mif.vu.lt

M. Vaičiulis
Vilnius university, Institute of Mathematics and Informatics, Lithuania
marijus.vaiciulis@mii.vu.lt

ABSTRACT

In the paper we review some new parametric family of functions which is used to construct tail index estimators. Functions from this new class are parameterized by two parameters are non-monotone in general, but presents a product of two monotone functions: the power function and the logarithmic function, which plays essential role in the classical tail index estimators. Many of the existing tail index (or more generally, extreme value) estimators can be expressed and some new estimators can be constructed using this family of functions. A unified approach to prove the consistency and asymptotic normality of estimators, based on this family of functions is provided. Introduced new estimators have better asymptotic performance comparing with the Hill estimator and other popular estimators over all range of the parameters present in the second order regular variation condition.

1. INTRODUCTION

In recent our papers two parametric families of functions, which are used to construct tail index estimators, were introduced. The main idea behind these families of functions is to include the logarithmic function, which plays an important role in

more general problem of extreme value index estimation, into some parametric family of functions. The first family of functions which was used in Paulauskas and Vaičiulis (2013) was

$$f_r(x) = \begin{cases} \frac{1}{r}\left(x^r - 1\right), & r \neq 0, \\ \ln x, & r = 0. \end{cases}$$

In the submitted paper Paulauskas and Vaičiulis (2014) more general family of functions was introduced

$$g_{r,u}(x) = x^r \ln^u(x),$$

and in both families functions are defined for $x \geq 1$. Let $X_{n,1} \leq X_{n,2} \leq \cdots \leq X_{n,n}$ denote the order statistic of $X_1, \ldots, X_n$, which are considered as i.i.d. random variables with distribution function F having regularly varying tail

$$\bar{F}(x) := 1 - F(x) = x^{-1/\gamma} L(x), \tag{1}$$

here L is a slowly varying at infinity function. Parameter γ is called extreme value index (EVI) and $\alpha := 1/\gamma > 0$ is called the tail index. Taking some part $k < n$ of the largest values from the sample we can construct the family of statistics

$$G_n(k, r, u) = \frac{1}{k} \sum_{i=0}^{k-1} g_{r,u}\left(\frac{X_{n-i,n}}{X_{n-k,n}}\right).$$

and taking various combinations of these statistics we can obtain many extreme value index γ estimators. In this way in Paulauskas and Vaičiulis (2013) there were obtained generalization of the Hill, the moment, and the moment ratio estimators (although in this paper functions f_r were used, but since there is simple relation between f_r and $g_{r,}u$, all estimators considered in this paper can be expressed via statistics $G_n(k, r, u)$). Under the so-called second order condition asymptotic normality of these generalized estimators was proved and it was shown that in the sense of asymptotic mean square error (AMSE) new estimators are better than classical ones. For example, it was shown that the generalized Hill estimator

$$\hat{\gamma}_n^{(1)}(k, r) = \frac{G_n(k, r, 0) - 1}{r \cdot G_n(k, r, 0)}$$

dominates (in the sense of AMSE) the classical Hill estimator

$$\hat{\gamma}_n^{(H)} = \hat{\gamma}_n^{(1)}(k,0) = G_n(k,0,1)$$

in the region of all possible values of the parameters $\gamma > 0, \rho < 0$, present in the second order condition (5). In Paulauskas and Vaičiulis (2014) the following generalization of the moment ratio estimator

$$\hat{\gamma}_n^{(2)}(k,r) = \begin{cases} (r \cdot G_n(k,r,1) - G_n(k,r,0) + 1)\,(r^2 \cdot G_n(k,r,1))^{-1}, & r \neq 0, \\ \gamma_n^{(MR)}(k), & r = 0, \end{cases}$$

was introduced and investigated. These two estimators $\hat{\gamma}_n^{(1)}(k,r)$ and $\hat{\gamma}_n^{(2)}(k,r)$ were the most successful among all generalizations, constructed in the papers Paulauskas and Vaičiulis (2013, 2014), in the sense that these estimators asymptotically dominates estimators, which they generalize, in all region of parameters present in the second order condition, and, most importantly, for these estimators there were obtained explicit (and rather simple) expressions for the optimal value of the parameter r. Also here it is appropriate to mention that the generalized Hill estimator $\hat{\gamma}_n^{(1)}(k,r)$ coincides with the so called Harmonic mean estimator, introduced in Beran *et al.* (2014) independently and at the same time as the paper Paulauskas and Vaičiulis (2013).

2. MAIN RESULTS

Let us denote the quantile function by

$$U(t) := F^{\leftarrow}(1 - 1/t) = \inf\{x : \ F(x) \geq 1 - 1/t\},$$

for $t \geq 1$, where $W \leftarrow : I \to \mathbb{R}$ is the left continuous inverse function of a monotone function W , defined by the relation $W \leftarrow(t) = \sup\{x : W(x) \leq t\}$, $t \in I$. The following result is well-known in extreme value theory (see, for example, De Haan and Ferreira, 2006): for $\gamma > 0$, condition (1) is equivalent to

$$\lim_{t\to\infty}\frac{U(tx)}{U(t)}=x^{\gamma}, \tag{2}$$

for all $x>0$, i.e., $U(t)\in RV\gamma$.

In order to construct the tail index estimators using the introduced statistics $G_n(k, r, u)$ we must know what quantities are estimated by these statistics, and our first result gives us the answer.

Theorem 2.1. Suppose that $X_1, \dots, X_n$ are i.i.d. nonnegative random variables whose quantile function U satisfies condition (2). Let $\gamma r<1$ and $u>-1$. Let us suppose that a sequence $k=k(n)$ satisfies conditions

$$k(n)\to\infty,\quad n/k(n)\to\infty,\quad \text{as } n\to\infty. \tag{3}$$

Then we have

$$G_n(k,r,u)\xrightarrow{\mathrm{P}}\xi(r,u):=\frac{\gamma^u\Gamma(1+u)}{(1-\gamma r)^{1+u}},\quad \text{as } n\to\infty.$$

Here $\xrightarrow{\mathrm{P}}$ stands for the convergence in probability and $\Gamma(u)$ denotes Euler's gamma function.

The following corollary allows to proof the consistency of estimators, expressed as a function of statistics $G_n(k, r, u)$ with different r and u.

Corollary 2.1. Suppose that $X_1, \dots, X_n$ are i.i.d. nonnegative random variables whose quantile function U satisfies condition (2). Let $\gamma r_j<1$ and $u_j>-1$, $j=1, 2, \dots, \ell$. Let us suppose that a sequence $k=k(n)$ satisfies (3). Let $\chi(t_1, \dots, t_l):(0,\infty)^{\ell}\to(0,\infty)$ be a continuous function. Then

$$\chi\left(G_n(k,r_1,u_1),\dots,G_n(k,r_\ell,u_\ell)\right)\xrightarrow{\mathrm{P}}\chi\left(\xi(r_1,u_1),\dots,\xi(r_\ell,u_\ell)\right), \tag{4}$$

as $n\to\infty$.

The relation (4) gives us many possibilities to form consistent estimators of γ using statistics $G_n(k, r, u)$ with different r, u, but since it is difficult to hope that it will be

possible to choose in optimal way two additional parameters r and u, we restricted ourselves with statistics $G_n(k, r, u)$ with $u = 0$ and $u = 1$. We note that only these statistics are in the expression of estimators $\hat{\gamma}_n^{(1)}(k, r)$ and $\hat{\gamma}_n^{(2)}(k, r)$. In order to compare these two estimators (which are, in a sense, the best among generalizations proposed in Paulauskas and Vaičiulis (2013, 2014)) with the estimators which do not involve additional parameter r, we had chosen the following two estimators:

$$\gamma_n^{(3)}(k) = \sqrt{\frac{G_n(k, 0, 2)}{2}}$$

and

$$\gamma_n^{(4)}(k) = \frac{-2 + G_n(k, 0, 1) + 3G_n(k, -1, 0) - G_n^2(k, -1, 0)}{1 - 2G_n(k, -1, 0) + G_n^2(k, -1, 0)}.$$

The estimators $\gamma_n^{(3)}(k)$ and $\gamma_n^{(4)}(k)$ were introduced in Gomes *et al.* (2000) and Fraga Alves *et al.* (2009), respectively. To prove the asymptotic normality of the introduced estimators we assume that there exists a measurable function $A(t)$ with the constant sign near infinity, which is not identically zero, and $A(t) \to 0$ as $t \to \infty$, such that

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho} \tag{5}$$

for all $x > 0$. Here $\rho < 0$ is the so-called second order parameter. It is known that (5) implies that the function $|A(t)|$ varies regularly with index ρ. Let us denote $d_r(k) = 1 - k\gamma r$.

Theorem 2.2. Suppose that $X_1, \ldots, X_n$ are i.i.d. nonnegative random variables whose quantile function U satisfies condition (5). Suppose that $\gamma r < 1/2$ and that the sequence $k = k(n)$ satisfies (3) and

$$\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = \mu \in (-\infty, +\infty). \tag{6}$$

Then, as $n \to \infty$,

$$\sqrt{k}\left(G_n(k, r, 0) - \xi(r, 0), G_n(k, r, 1) - \xi(r, 1)\right) \xrightarrow{\mathrm{d}} \mu\left(\nu^{(1)}(r), \nu^{(2)}(r)\right) + \left(W^{(1)}, W^{(2)}\right)$$

where $\xrightarrow{\mathrm{P}}$ stands for the convergence in distribution and quantities $\nu^{(j)}(r), j = 1, 2$ are as follows

$$\nu^{(1)}(r) = \frac{r}{d_r(1)(d_r(1)-\rho)}, \quad \nu^{(2)}(r) = \frac{1-\rho-\gamma^2 r^2}{(d_r(1))^2(d_r(1)-\rho)^2}.$$

Here $(W^{(1)}, W^{(2)})$ is zero mean Gaussian random vector with variances $\mathrm{E}\left(W^{(j)}\right)^2 = s_j^2(r), j = 1, 2$ and covariance $E(W^{(1)}, W^{(2)}) = s_{12}(r)$, where

$$s_1^2(r) = \frac{\gamma^2 r^2}{d_r(2)d_r^2(1)}, \quad s_2^2(r) = \frac{\gamma^2(d_r(2)+2\gamma^4 r^4)}{d_r^3(2)d_r^4(1)}, \quad s_{12}(r) = \frac{\gamma^2 r(d_r(1)-\gamma^2 r^2)}{d_r^2(2)d_r^3(1)}.$$

Proofs of Theorems 2.1 and 2.2 can be found in Paulauskas and Vaičiulis (2014). This result gives us the general method (application of Continuous Mapping Theorem) to prove the asymptotic normality of many estimators, which are expressed by statistics $G_n(k, r, u)$ with $u = 0$ and $u = 1$ (it is possible to formulate analogous result with the pair of arbitrary u_1 and u_2, only expressions of limit parameters will be of more complicated form). We provide the following theorem.

Theorem 2.3. Suppose that $X_1, \ldots, X_n$ are i.i.d. nonnegative random variables whose quantile function U satisfies condition (5). Suppose that the sequence $k = k(n)$ satisfies (3) and (6). Then, as $n \to \infty$,

$$\sqrt{k}\left(\hat{\gamma}_n^{(i)}(k) - \gamma\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mu\nu_i, \sigma_i^2\right), \quad i = 3, 4,$$

where $\xrightarrow{\mathrm{d}}$ stands for the convergence in distribution. If, in addition, $\gamma r < 1/2$, then, as $n \to \infty$,

$$\sqrt{k}\left(\hat{\gamma}_n^{(i)}(k, r) - \gamma\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mu\nu_i(r), \sigma_i^2(r)\right), \quad i = 1, 2.$$

The asymptotic quantities $\nu_i(r)$, $\sigma_i^2(r)$, $i = 1, 2$ and ν_i, σ_i^2, $i = 3, 4$ are as follows

$$\begin{aligned} \nu_1(r) &= \frac{d_r(1)}{d_r(1)-\rho}, \quad \sigma_1^2(r) = \frac{\gamma^2 d_r^2(1)}{d_r(2)}, \\ \nu_2(r) &= \frac{d_r^2(1)}{(d_r(1)-\rho)^2}, \quad \sigma_3^2(r) = \frac{2\gamma^2 d_r^4(1)}{d_r^3(2)}, \end{aligned}$$

$$\begin{aligned}\nu_3 &= \frac{2-\rho}{2(1-\rho)^2}, \quad \sigma_3^2 = \frac{5\gamma^2}{4}, \\ \nu_4 &= \frac{(1+\gamma)(\gamma+\rho)}{\gamma(1-\rho)(1-\rho+\gamma)}, \quad \sigma_4^2 = (1+\gamma)^2.\end{aligned}$$

Proof of this result in the case $i = 3$ is in Gomes *et al.* (2000), the case $i = 4$ is considered in Fraga Alves *et al.* (2009), while the cases $i = 1, 2$ can be found in Paulauskas and Vaičiulis (2014). This result is used to study AMSE of these estimators and to compare them, this is done in Paulauskas and Vaičiulis (2014). In the paper Paulauskas and Vaičiulis (2014a) the behavior of these estimators for the middle size samples is investigated.

ACKNOWLEDGMENTS

The research was supported by Research Council of Lithuania, grant No. MIP-076/2013.

REFERENCES

Beran, J., Schell D., and Stehlík, M. (2014). "The Harmonic Moment Tail Index Estimator: Asymptotic Distribution and Robustness". *Ann. Inst. Stat. Math.*, 66, 193-220.

De Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New York.

Fraga Alves, M.I., Gomes, M.I., de Haan, L., and Neves, C. (2009). "Mixed Moment Estimator and Location Invariant Alternatives". *Extremes*, 12, 149-185.

Gomes, M.I., Martins, M.J., and Neves, M. (2000). "Alternatives to a Semiparametric Estimation of Parameters of Rare Events – The Jacknife Methodology". *Extremes*, 3, 207-229.

Paulauskas, V., and Vaičiulis, M. (2013). "On the Improvement of Hill and some other Estimators". *Lithuanian Math. J.*, 53, 336-355.

Paulauskas, V., and Vaičiulis, M. (2014). *Several New Tail Index Estimators* (submitted for publication).

Paulauskas, V., and Vaičiulis, M. (2014a). *Comparison of Several Tail Index Estimators* (in preparation).

71. FLEXIBLE BONUS–MALUS SYSTEMS INDUCED BY OWA OPTIMIZATION

V. Blanco-Izquierdo
University of Granada, Department of Quantitative Methods, Spain
vblanco@ugr.es

J.M. Pérez-Sánchez
University of Granada, Department of Quantitative Methods, Spain
josemag@ugr.es

ABSTRACT

Ordered weighted averaging (OWA) operators were introduced by Yager (1988) as a flexible approach to aggregate functions and that allows to model as particular cases the maximum, the mean, or the minimum of a finite set of functions. In OWA aggregation, for a given vector of weights, they are assigned to ordered values, the first weight to the largest value, the second weight to the second largest value, etc. In this paper we use OWA operators to compute risk premiums in problems typically arising in a Bonus-Malus System (BMS).

1. INTRODUCTION

Ordered weighted averaging (OWA) operators were introduced by Yager (1988) as a flexible approach to aggregate functions and that allows to model as particular cases the maximum, the mean, or the minumum of a finite set of functions. In OWA aggregation, for a given vector of weights, they are assigned to ordered values, the first weight to the largest value, the second weight to the second largest value, etc.

Hence, in order to aggregate n functions $f_1(x), \ldots, f_n(x) : R^n \to R$, first for each x, the vector $\mathbf{f}(x) = f_1(x), \ldots, f_n(x)$ is sorted in non-decreasing order, obtaining $(f_{\sigma(1)}(x), \ldots, f_{\sigma(n)}(x))$, for a permutation σ on the indices $\{1, \ldots, n\}$, such that $f\sigma(1)(x) \geq \cdots \geq f\sigma(n)(x)$. Then, for the given set of weights $\omega_1, \ldots, \omega n \in R_+$, the OWA aggregation function is defined as:

$$\mathrm{OWA}(\mathbf{f}; \omega_1, \ldots, \omega_n) := \sum_{i=1}^{n} \omega_i \; f_{\sigma i}(x)$$

where the i-th weight, ω_i is assigned to the value $f_{\sigma i}(x)$ which is in the i-th position when the list $\mathbf{f}(x)$ is sorted in non-decreasing order. Note that, in case $\omega_i = 1$ for all i, the OWA operator coincides with the mean operator, while if $\omega_1 = 1$ and $\omega_i = 0$ for $i \neq 1$, it coincides with the maximum operator. OWA aggregation has been applied in many different fields in order to generalize classical operators. In particular, OWA operators have been used to flexibilize standard objective function in linear regression models, location theory. See Nickel and Puerto (2005), among others. In this paper we use OWA operators to compute risk premiums in problems typically arising in a Bonus-Malus System (BMS). BMS is a merit rating method that is widely used in automobile insurance, characterized by the fact that only the number of claims modifies the premium. The methodology of BMS consists of ensuring that the premium increases with the number of claims and decreases with the period t in which the policyholder does not make a claim. This can be achieved by dividing a posterior expectation by a prior expectation according to an estimate derived by the expected means of an appropriate loss function (Lemaire, 1979). The proposed approach is used to compute Bayesian premiums, which are widely used in setting automobile third-party liability insurance in the context of BMS (see Gómez *et al.*, 2002, 2006; Lemaire, 1979, 1988, 1998 and Meng and Whitmore, 1999; among others). We provide mathematical programming formulations for the problems, whose optimal solutions are the premium that should be applied to a certain set of policyholders. The main differences of our approach with respect to the previous one appeared in the literature are: (1) The premium assigned to a client depends not only of his behaviour (in terms of his particular claims) but also of the claims of the rest of the clients; (2) the distribution of clients with respect of the number of claims is considered; and (3) the problem is considered as a flexible optimization problem whose parameters may be modified conveniently. A model often used for experience rating in a BMS assumes that each individual risk has its own Poisson distribution

for a number of claims, assuming that the mean number of claims is distributed across individual policyholders according to a Gamma distribution, thus taking advantage of its good analytical properties for conjugance (Coene and Doray, 1996; Corlier *et al.*, 1979; Dionne and Vanasse, 1989; Lemaire, 1979, 1988, 1998, among others). Other possibilities have been studied previously in BMS settings. For example, Tremblay (1992) analized the Poisson-inverse Gaussian model in order to fit automobile insurance data and compute risk premiums. Walhin and Paris (1999) considered a finite Poisson mixture and Hofmann's distributions as claim frequency distributions. Meng and Whitmore (1999) considered the negative binomial-Pareto distribution for the same purposes. Dionne and Vanasse (1989) and, more recently, Mahmoudvand and Hassani (2009) generalized the analysis by introducing the severity component in the premiums calculation.

2. OWA BONUS–MALUS SYSTEM

In this section, we propose flexible approaches to deal with Bonus–Malus premiums ratings. As particular cases we will obtain the Lemaire approach. Through this section, instead of computing premiums for single customers, we consider the whole portfolio, in such a way that the premium of a client (or a group of similar clients) may affect the premium of a different client. Let $C = \{C_1, \ldots, C_n\}$ a set of clients and $T = \{0, \ldots, t\}$ the time period. Our goal is to assign assign to each group of clients C_i and each period k, the premium P_{ik}, for $i = 1, \ldots, n$ and $k = 0, \ldots, t$. Such an assignment is performed to minimize the overall loss produced by the clients. We assume that the losses are quadratic. In the classical theory, for each group of clients C_i and each period k, the (risk, collective or bayes) premium is minimized to get the optimum premium. For the sake of readability, we denote the three different types of objective functions (risks), for the net premium principle and Poisson-Gamma distributions as:

$$\mathrm{RPO}(P, x; \lambda) = \sum_{\lambda=0}^{\infty} (P - x)^2 f_\lambda(x) dx = P^2 - 2\lambda P + \lambda(1 + \lambda)$$

$$\mathrm{CPO}(P; \lambda, \alpha, \beta) = \int_0^\infty (P - \lambda)^2 \pi_0(\lambda) d\lambda = P^2 - 2\frac{\alpha}{\beta} P + \frac{\alpha(\alpha + 1)}{\beta}$$

$$\mathrm{BPO}(P; \alpha, \beta; (k, t)) = P^2 - 2\frac{\alpha + k}{\beta + t} P + \frac{(\alpha + k)(\alpha + k + 1)}{(\beta + t)^2}$$

Note that, under our assumptions, the three functions are quadratic on $P \in R_+$. Our goal is to provide a general methodology to compute the risk, collective and Bayes premiums for a whole portfolio of client. Let $\{p_1, \ldots, p_n\} \subset R_+$ and such that $\Sigma i\, p_i = 1$, where p_i is the proportion of clients in C_i. Also, consider a set of weights $\omega = (\omega_1, \ldots, \omega_n] \in R_+$.

Definition 1 (OWA Risk Premium). Let $\lambda_1, \ldots, \lambda_n$ the mean number of claims for each of the sets in the portfolio C. Let also $\mathrm{F}(\mathbf{P}, \mathrm{x}; \lambda) : R^{\mathrm{n}} \times R^{\mathrm{n}} \to R^{\mathrm{n}}_{+}$ be defined as $F_i(\mathrm{x}; \lambda) = \mathrm{RPO}(x_i; \lambda_i)$. Then, the risk premiums for the portfolio, $\mathbf{P} = (P_1, \ldots, P_n)$, are obtained by minimizing the following OWA function: $\sum_{i=1}^{n} \omega_i\, p_{(i)} \mathrm{F}_{(i)}(\mathbf{P}, \lambda)$

Note that the ith omega weight is assigned to the risk premium value which is sorted in the ith position, while the proportion p_i is always assigned to the risk premium for the ith group (the one corresponding with such a percentage of clients).

In the following result we provide a mathematical programming formulation for the computation of the above OWA risk net premium.

Theorem 2. Let $C = \{C_1, \ldots, C_n\}$ be a set of clients, and let λ_j be the mean number of claims of group C_j, for $j = 1, \ldots, n$. Then, the OWA risk premiums $P_1, \ldots, P_n$ are the optimal solutions of the following problem.

$$
\begin{aligned}
&\min \sum_{k=1}^{n} \omega_k \theta_k \\
s.t. &\sum_{i=1}^{n} z_{ki} \leq k - 1, && \forall k = 1, \ldots, n, \\
&Q_i \geq p_i \,\mathrm{RPO}(P_i, \lambda_i), && \forall i = 1, \ldots, n, \\
&Q_i \leq \theta_k + M\, z_{ki}, && \forall i, k = 1, \ldots, n, \qquad \text{(OWARP)} \\
&\theta_k \geq \theta k + 1, && \forall k = 1, \ldots, n-1, \\
&Q_i, P_i, \theta_i \geq 0, && \forall i = 1, \ldots, n, \\
&z_{ik} \in \{0, 1\}, && \forall i, k = 1, \ldots, n.
\end{aligned}
$$

In what follows, we determine a similar method to extend OWA risk premiums to OWA collective premiums. For the sake of that, for the each of the groups in the portfolio of clients $C = \{C_1, \ldots, C_n\}$ we consider that the average number of claims

(which is actually an unknown parameter) for C_j is distributed by a Gamma random variable with parameters α_j and β_j, for $j = 1, \ldots, n$. Let $\{p_1, \ldots, p_n\}$ the proportions of clients on each of the groups and the OWA aggregation weights $\{\omega_1, \ldots, \omega_n\}$.

Definition 3 (OWA Collective Premium). Let $G(\mathrm{x}; \alpha, \beta) : R_n \to R^{\mathrm{n}}_{+}$ be defined as $\mathrm{G}_i(\mathrm{x}; \alpha, \beta) = \mathrm{CPO}(x_i; \alpha_i, \beta_i)$. Then, the collective premiums for the portfolio, $\mathbf{P} = (P_1, \ldots, P_n)$, are obtained by minimizing the following OWA function: $\sum_{i=1}^{n} \omega_i \, p_{(i)} \mathrm{G}_{(i)}(\mathbf{P}; \alpha, \beta)$.

Analogously to the risk premium, the OWA collective premiums for the portfolio C can be computed as the optimal solutions of an easy-to-handle mathematical programming problem.

Theorem 4. Let $C = \{C_1, \ldots, C_n\}$ be a set of clients. The OWA collective premiums $P_1, \ldots, P_n$ are the optimal solutions of the following problem.

$$
\begin{array}{lll}
\min \displaystyle\sum_{k=1}^{n} \omega_k \theta_k & & \\
s.t. \displaystyle\sum_{i=1}^{n} z_{ki} \leq k - 1, & \forall k = 1, \ldots, n, & \\
Q_i \geq p_i \, \mathrm{CPO}(P_i; \alpha_i, \beta_i), & \forall i = 1, \ldots, n, & \\
Q_i \leq \theta_k + M \, z_{ki}, & \forall i, k = 1, \ldots, n, & \text{(OWACP)} \\
\theta_k \geq \theta k + 1, & \forall k = 1, \ldots, n-1, & \\
Q_i, P_i, \theta_i \geq 0, & \forall i = 1, \ldots, n, & \\
z_{ik} \in \{0, 1\}, & \forall i, k = 1, \ldots, n. &
\end{array}
$$

Finally, we define the OWA Bayes premium as follows.

Definition 5 (OWA Bayes Premium). Let $\mathrm{H}(\mathrm{x}; \alpha, \beta, (k, t)) : R_{\mathrm{n}} \to R^{\mathrm{n}}_{+}$ be defined as $\mathrm{H}_i(x; \alpha, \beta, (k, t)) = BPO(x_i; \alpha_i, \beta_i, (k, t))$, for a given number of claims k and a fixed period t. Then, the Bayes premiums for the portfolio, $\mathbf{P} = (P_1, \ldots, P_n)$, for those clients with k claims in period t are obtained by minimizing the following OWA function: $\sum_{i=1}^{n} \omega_i \, p_{(i)} \mathrm{H}_{(i)}(\mathbf{P}; \alpha, \beta, (k, t))$.

3. EXPERIMENTS

In this section, we apply all the previous results into an insurance data set of policies. This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005 in Australia. There are 67,856 policies, of which 4,624 (6.8%) had at least one claim. The data set contains the number of claims of each policyholder during the considered period. Table 1 shows the new OWA Bonus-Malus premiums and table 2 these premiums calculated under standard Lemaire's modelization in order to compare both.

C1	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	95.08	156.66	217.28	278.25	340.67
t = 2	90.62	147.21	207.20	265.44	325.39
t = 3	86.55	140.68	196.26	253.73	308.09
C2	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	93.98	179.51	265.06	350.84	437.78
t = 2	88.67	168.04	250.23	329.41	409.63
t = 3	83.91	159.49	234.71	310.66	386.56
C3	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	93.31	196.18	298.03	402.47	502.12
t = 2	87.96	184.97	281.07	377.49	472.94
t = 3	83.19	174.39	267.11	357.94	449.53

Table 1. BMP under OWA optimitation.

C1	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	94.93	156.75	218.57	280.39	342.22
t = 2	90.36	149.20	208.04	266.88	325.72
t = 3	86.20	142.34	198.47	254.61	310.75
C2	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	93.62	179.43	265.23	351.04	436.85
t = 2	88.01	168.68	249.34	330.01	410.67
t = 3	83.04	159.14	235.24	311.35	387.45
C3	k = 0	k = 1	k = 2	k = 3	k = 4
t = 0	100				
t = 1	93.88	195.77	297.65	399.54	501.43
t = 2	88.46	184.48	280.49	376.51	472.52
t = 3	83.64	174.42	265.20	355.97	446.75

Table 2. BMP under Lemairet's ratio.

REFERENCES

Dionne, G., and Vanasse, C., (1989). "A Generalization of Actuarial Automobile Insurance Rating Models: The Negative Binomial Distribution with a Regression Component". *ASTIN Bulletin*, 19 (2), 199-212.

Gómez, E., Hernández, A., Pérez, J.M., and Vázquez-Polo, F.J. (2002). "Measuring Sensitivity in a Bonus-Malus System". *Insurance: Mathematics & Economics*, 31, 105-113.

Gómez-Déniz, E., Vázquez-Polo, F.J., and Pérez, J.M. (2006). "A Note on Computing Bonus-Malus Insurance Premiums Using a Hierarchical Bayesian Framework". *Test*, 15 (2), 345-359.

Lemaire, J. (1979). "How to Define a Bonus-Malus System with an Exponential Utility Function". *ASTIN Bulletin*, 10 (3), 274-282.

Lemaire, J. (1988). "Construction of the New Belgian Motor Third Party Tariff Structure". *ASTIN Bulletin*, 18 (1), 99-112.

Lemaire, J. (1998). "Bonus-Malus System: The European and Asian Approach to Merir-Rating (with Discussion)". *North American Actuarial Journal*, 2 (1), 1-22.

Mahmoudvand, R., and Hassani, H. (2009). "Generalized Bonus-Malus Systems with a Frequency and a Severity Component on an Individual Basis in Automobile Insurance". *ASTIN Bulletin*, 39 (1), 307-315.

Meng, Y.W., and Whitmore, G.A. (1999). "Accounting for Individual Over-Dispersion in a Bonus-Malus Automobile Insurance System". *ASTIN Bulletin*, 29 (2), 327-337.

Nickel, S., and Puerto, J. (2005). *Location Theory: A Unified Approach*. Berlin, Springer.

Tremblay, L. (1992). "Using the Poisson Inverse Gaussian in Bonus-Malus Systems". *ASTIN Bulletin*, 22 (1), 97-106.

Walhin, J.F., and Paris, J. (1999). "Using Mixed Poisson Processes in Connection with Bonus-Malus Systems". *ASTIN Bulletin*, 29 (1), 81-99.

Yager, R.R. (1988). "On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making". *IEEE Trans. Sys., Man and Cyber.*, 18, 183-190.

72. THE EFFECT OF ZERO-INFLATION IN CLAIMS INSURANCE THROUGH A BAYESIAN MODEL

J.M. Pérez-Sánchez
University of Granada, Quantitative Methods, Spain
josemag@ugr.es

E. Gómez-Déniz
University of Las Palmas de Gran Canaria and TiDES Institute, Quantitative Methods, Spain
emilio.gomez-deniz@ulpgc.es

ABSTRACT

In claims count data, there is usually an excessive number of observations with a count of zero and Poisson regression model, the most used in the actuarial literature, often fits this kind of data poorly, as indicated by a deviance or Pearson chi–square test. This is so because for the Poisson distribution the conditional variance of the dependent variable is equal to the conditional mean. Then the data shows overdispersion phenomena. Zero-inflated models seem the best way to model this setting. The zero-inflated Poisson (ZIP) model is one way to allow for overdispersion. This model assumes two groups: one whose counts are generated by the standard Poisson regression model, and another group who have zero probability of a count greater than zero.

1. INTRODUCTION

Poisson regression models provide a standard framework for the analysis of count data. However, count data show often overdispersion (variance larger than the mean) relative to the Poisson distribution. One frequent factor of overdispersion is

that the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special issue. For example, in counting claims from policyholders, a policyholder may have no claims either because he/she is a good driver, or simply because no risk factors have happened "near" his/her driving. This is the distinction between structural zeros, which are (almost) inevitable, and sampling zeros, which occur by chance.

In recent years there has been considerable interest in models for count data that allow for excess zeros, particularly in the econometric literature. Mullahy (1986) explores the specification and testing of some modified count data models. Lambert (1992) provides a manufacturing defects application of these models and discusses the case of zero-inflated Poisson (ZIP) models. Gupta *et al*. (1996) provide a general analysis of zero-inflated models. Gurmu (1997) develops a semi-parametric estimation method for hurdle (two-part) count regression models. Ridout *et al*. (1998) consider the problem of modelling count data with excess zeros and review some possible models. Hall (2000) adapt Lambert's methodology to an upper bounded count situation, thereby obtaining a zero–inflated binomial (ZIP) model. Ghosh *et al*. (2003) introduce a flexible class of zero inflated models which includes other familiar models such as the Zero Inflated Poisson (ZIP) models, as special cases by using a Bayesian estimation method. An overview of count data in econometrics including zero inflated models is provided in Cameron and Trivedi (1998, 2005). Application areas are diverse and have included health sciences, medicine, ecology, econometric, finance, actuarial and road safety. In insurance, Yip and Yao (2005) provide a better fit to their insurance data by using zero-inflated count models and Boucher *et al*. (2007) compare zero-inflated and hurdle models in an automobile insurance portfolio. Recently, Mouatassim and Ezzahid (2012) analyze the zero-inflated models with an application to private health insurance data.

In this paper we review and compare several inflated and standard models with particular focus on applications in claims count data. Section 2 provides a survey of models that have been analyzed. Sections 3 looks at insurance application and 4 briefly discuss the results.

2. MODELS FOR CLAIM DATA

In this section a survey of models that will be used in this work are briefly introduced.

2.1. Standard models

The Poisson (log-linear) regression model is the most basic model that explicitly takes into account the nonnegative integer–valued aspect of the dependent count variable. In this model, the probability of a claim count yi, given the vector of covariates x_i is modelled by a Poisson distribution with parameter $\lambda > 0$. Then,

$$P(Y_i = y_i|x_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ... \tag{1}$$

The mean parameter λ (the conditional mean number of claims) is a function of the vector of covariates:

$$E(y_i|x_i) = \lambda = \exp(x_i'\beta), \tag{2}$$

where β is a $(k+1) \times 1$ parameter vector. The name log–linear model is also used for the Poisson regression model because the logarithm of the conditional mean is linear in the parameters:

$$\ln[E(y_i|x_i)] = \ln(\lambda) = x_i'\beta. \tag{3}$$

The Poisson regression model assumes that the data are equally dispersed, that is, that the conditional variance equals the conditional mean. The Poisson model has been criticized for this restrictive property. Real–life data are often characterized by overdispersion, that is, the variance exceeds the mean. The negative binomial regression model is a generalization of the Poisson regression model that allows for overdispersion by introducing an unobserved heterogeneity term for observation i. Observations are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. In the negative binomial model,

$$E(y_i|x_i) = \lambda\mu_i = \exp(x_i'\beta)\mu_i, \tag{4}$$

where μ_i follows a gamma$(\theta; \theta)$ distribution with $E(\mu_i) = 1$ and $var(\mu_i) = 1/\theta$. Conditional on x_i, dependent count variable Y_i is distributed as a neg ative binomial:

$$P(Y_i = y_i|x_i) = \frac{\theta^\theta \lambda^{y_i} \Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)(\lambda + \theta)^{\theta+y_i}}. \tag{5}$$

The conditional mean and variance are

$$E(y_i|x_i) = var(y_i|x_i) = \lambda(1 + (1/\theta)\lambda). \tag{6}$$

2.2. Zero-inflated models

Zero–inflated Poisson (ZIP) distribution is given by

$$Pr(Y_i = y_i) = \begin{cases} w + (1-w)\exp(-\lambda), & y_i = 0 \\ (1-w)\exp(-\lambda)\lambda^{y_i}/y_i! & y_i > 0 \end{cases}, \tag{7}$$

where the outcome variable yi has any non–negative integer value, λ is the expected Poisson count and w is the probability of extra zeros. The first process is governed by a binary distribution that generates structural zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero. The mean and the variance are given by

$$E(y_i) = (1-w)\lambda, \qquad var(y_i) = (1-w)\lambda(1 + w\lambda). \tag{8}$$

Zero-inflated forms of other count distributions, such as the negative binomial, can be defined similarly. Gupta *et al*. (1996) and Ghosh *et al*. (2006), for example, investigated the zero-inflated form of the generalized Poisson distribution. Ghosh *et al*. (2006) considered that the zero–inflated model can be represented as $Y = V(1 - B)$, where B is a Bernoulli(p) random variable and V independently to B has a discrete distribution such as Poisson(θ), NegBin$(\theta,$ r$)$ or more generally power series, PS(θ). Under this representation, the mean $(E(Y))$ and variance $(var(Y))$ are given by $E(y) = (1 - w)E(V)$ *and* $V(y) = w/(1 - w)E(y)^2 + \delta E(y)$, where $\delta = var(V)/E(V)$ denotes the coefficient of dispersion of the latent random variable V.

Variables	Mean	Variance	Minimum	Maximum
Number of claims	0.07275	0.07739	0	4
Vehicle value	1.77702	1.45258	0	34.56
Gender	0.43110	0.24525	0	1
Young age	0.27436	0.19908	0	1
Medium age	0.57492	0.24439	0	1
Old age	0.46081	0.24846	0	1
Vehicle age	0.57492	0.24439	0	1

Table 1. Descriptive summary of variables.

Thus, it follows that if the latent variable, V does not have an underdispersed distribution (i.e., $\delta \geq 1$), then the distribution of Y is overdispersed. Onthe other hand, if V has a underdispersed distribution (i.e., $\delta < 1$), then Y has a underdispersed distribution if and only if $E(V) < (1 - \delta)/w$. In their paper, they suppose that V follows the power series (PS) distribution. So, the probability mass function of their $ZIPS(\mathrm{w}, \theta)$ distribution is given by,

$$Pr(Y_i = y_i) = \begin{cases} w + (1-w)\dfrac{b(0)}{c(\theta)}, & y_i = 0 \\ (1-w)\dfrac{b(k)\theta^k}{c(\theta)} & y_i = k, \quad k = 1, 2, \ldots \end{cases}, \tag{9}$$

where $c(\theta) = \sum_{k=0}^{\infty} b(k)\theta^k$, $0 \leq w < 1$ and $\theta > 0$. The mean is $E(Y) = (1 - w)\mu(\theta)$ where $\mu(\theta)$ denotes the mean of a regular $PS(\theta)$ distribution. The covariates are usually linked to model parameters w and θ in this way:

$$\log(\mu) = Z\beta; \qquad \text{logit}[w/(1-w)] = P\gamma, \tag{10}$$

with β and γ vectors of regression parameters associated with covariates $Z = (z_1, \ldots, z_n)'$ and $P = (p_1, \ldots, p_w)'$, respectively. As they comment, it is possible to suppose $Z = P$ in some applications.

3. APPLICATION TO INSURANCE DATA

The data studied contain information from policyholders of several Australian insurance companies in 2004 or 2005, describing certain characteristics related to the

the vehicles and the policyholders. There are 67,856 policies, of which 63,232 (93.18%) had no claims. Table 1 shows a descriptive of the dependent and independent variables. For each policy, the initial information for the period considered and the existence or otherwise of at least a claim were reported within this yearly period. In total, four explanatory variables are considered, together with a dependent variable representing the number of claims. Vehicle value represents its value in 10,000 Australian dollars. Vehicle age is equal to 1 if the vehicle is relatively young (7 years or less). Gender is equal to 1 if the policyholder is a woman. Finally, a categorical variable was considered to represent the age of the policy-holder by dividing this feature in three dummies: young, medium and old ages. Table 2 shows standard results under Poisson, Negative binomial, zero-inflated Poisson and zero-inflated Negative binomial regressions. As we can see, vehicle value and youngest policyholders (in relation to medium age) are positively significant for the number of claims. In the other way, oldest people are expected to decreased their number of claims in relation to medium age policyholders. At 10 % level of significance, vehicle age is also negatively significant.

	Poisson	Negative Binomial	ZIP	ZIBN
Vehicle value	0.0381 ***	0.0392 ***	0.0392 ***	0.0393 ***
	(0.0114)	(0.0117)	(0.0118)	(0.0117)
Gender	-0.0159	-0.0161	-0.0164	-0.0162
	(0.0300)	(0.0300)	(0.0301)	(0.0301)
Young age	0.0958 ***	0.0955 ***	0.0949 ***	0.0956 ***
	(0.0337)	(0.0337)	(0.0337)	(0.0337)
Old age	-0.2081 ***	-0.2083 ***	-0.2082 ***	-0.2083 ***
	(0.0385)	(0.0385)	(0.0385)	(0.0385)
Vehicle age	-0.0617 *	-0.0606 *	-0.0604 *	-0.0606 *
	(0.0334)	(0.0335)	(0.0335)	(0.0335)
Inflation constant			-0.2849 *	-14.1262 ***
			(0.1283)	(0.7096)
Deviance	36118.353	36019.507	36026.881	36019.507
AIC	36130.353	36033.507	36040.881	36035.507

Table 2. Standard fitted models.

	ZIPS (α_i)	ZIPS (β_i)
Vehicle value	0.5291 ***	-0.0955 ***
	(0.102)	(0.0297)
Gender	0.0022	-0.0275
	(0.1985)	(0.0831)
Young age	-0.0743	0.1225
	(0.102)	(0.0826)
Old age	-0.5273 ***	0.0261
	(0.2072)	(0.0962)
Vehicle age	0.0111	-0.0118
	(0.2565)	(0.0982)
Deviance		9617.0
AIC		9629.0

Table 3. Zero inflated Power Series fitted model.

Table 3 shows the results under Zero-inflated Power series model. As we can see, this model fits better than the other ones (lower desviance and AIC values).

4. CONCLUSION

In this paper, we have reviewed standard and zero-inflated count models and apply them to modelling actuarial claim data. We have analyzed which relevant factors detects the models and we have shown that Power series model under Bayesian consideration got the best fits.

REFERENCES

Boucher, J.-P., Denuit, M., and Guillén, M. (2007). "Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models". *North American Actuarial Journal*, 11, 110-131.

Cameron, A.C., and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Cameron, A.C., and Trivedi, P.K. (2005). *Microenometrics: Methods and Applications*. Cambridge University Press.

Ghosh, S.K., Mukhopadhyay, P., and Lu, J.-C. (2003). "Bayesian Analysis of Zero-Inflated Regression Models". *Journal of Statistical Planning and Inference*, 136, 1360-1375.

Gupta, P.L., Gupta, R.C., and Tripathi, R.C. (1996). "Analysis of Zero-Adjusted Count Data". *Comput. Statist. Data Anal.*, 23, 207-218.

Gurmu, S. (1997). "Semi-Parametric Estimation of Hurdle Regression Models with an Application to Medicaid Utilization". *Journal of Applied Econometrics*, 12, 225-242.

Hall, D.B. (2000). "Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study". *Biometrics*, 56-4, 1030-1039.

Lambert, D. (1992). "Zero-inflated Poisson Regression, with an Application to Defects In Manufacturing". *Technometrics*, 34, 1-14.

Mouatassim, Y., and Ezzahid, E.H. (2012). "Poisson Regression and Zero-Inflated Poisson Regression: Application to Private Health Insurance Data". *European Actuarial Journal*, 2-2, 187-204.

Mullahy, J. (1986). "Specification and Testing of some Modified Count Data Models". *Journal of Econometrics*, 33, 341-365.

Ridout, M., Demétrio, C.G.B., and Hinde, J. (1998). "Models for Count Data with many Zeros". *International Biometrics Conference*. Cape Town, South Africa.

Yip, K., and Yau, K.K.W. (2005). "On Modeling Claim Frequency Data In General Insurance with Extra Zeros". *Insurance: Mathematics and Economics*, 36-2, 153-163.

73. FORECASTING PROBABILITY OF BANKRUPTCY FROM UNBALANCED DATA

F. Pierri
University of Perugia, Department of Economics, Statistical Section, Italy,
francesca.pierri@unipg.it

E. Stanghellini
University of Perugia, Department of Economics, Statistical Section, Italy,
elena.stanghellini@unipg.it

N. Bistoni
Graduate of Department of Economics, University of Perugia, Italy,
nicobistoni@yahoo.it

ABSTRACT

When analysing the determinants of bankruptcy of small and medium enterprises, one of the most common problems is that of unbalanced data, as very often the event under study happens in only a small percentage of cases. The aim of this paper is to explore three different statistical methods of coping with unbalanced data and to identify which of these has the greatest predictive capability in the context of the bankrupcty event.

The dataset is composed of all firms which were active in Tuscany in 2006. For each of them we have a five-year series of balance sheet indicators. Bankruptcy is represented by their legal status at May 2010. We focused on some indicators previously identified as predictors of the state of bankruptcy (Pierri, 2013; Pierri, Burchi and Stanghellini, 2013) and we tested the same model using the following three methods: logistic regression for matched case-control studies, logistic regression for a random balanced data sample, logistic regression for a sample balanced by ROSE

(Random OverSampling Examples, Menardi and Torelli, 2014). We built a training sample to develop the models and a hold-out sample to compare their discriminatory ability through ROC curves.

1. INTRODUCTION

The measurement of risk concerns classifying out-of-sample units into two categories: bad and good. A large variety of statistical models is available, known as credit scoring techniques, and logistic regression is widely used in this context. Frequently one of the two events is rare, so even in the case of all categorical explanatory variables, contingency tables will have very low or zero frequencies in the cells related to this event. Things get more extreme when there is at least one continuous variable in the set of the explanatory variables. In this situation, estimation of the logistic regression model may lead to high classification errors of rare units. The aim of this paper is to compare different techniques which allow accurate estimation under these conditions.

The study is carried out on real data, selected from the "AIDA" database, concerning balance sheet indicators of companies in the Tuscany region of Italy which contains a large number of small and medium-sized enterprises. The event of interest is the opening of insolvency proceedings for bankruptcy which, luckily from an economic point of view, can be considered rare.

In order to face this problem we applied logistic regression using different sampling techniques: case-control sampling, balanced random sampling and random oversampling (ROSE method). From the full dataset we built a training and a hold-out sample: the first one forms the basis of data for the implementation of the different methodologies, and the second is used to compare the three classification methods on the basis of the ROC Curve.

The theoretical illustration of the three methodologies is followed by a brief description of the data and their application. The three models are then compared.

2. THE LOGISTIC MODEL WITH BALANCED DATA

A first methodological choice to solve the problem of unbalanced data is the use of a retrospective model. Let Y be the Bernoulli random variable taking value 1 for success (good firm) and 0 otherwise (bankruptcy). Starting from the training sample, we consider all the observations with Y = 0 and a random sample of observations where Y = 1 in such a way that the marginal distribution of Y in the retrospective sample has 4 good cases for each bad one.

Let Z be a variable which takes the value 1 if a unit is included in the population from which the balanced sample is drawn and 0 otherwise; moreover define $K_1 = P(Z = 1 | Y = 1)$ the probability to extract a healthy unit and $K_0 = P(Z = 1 \mid Y = 0)$ the complementary probability, both independent of a p dimensional vector x of covariates. Assuming the population model is logistic, on the basis of Bayes theorem we can write:

$$log \frac{P(Y=1 \mid \mathbf{x}, Z=1)}{P(Y=0 \mid \mathbf{x}, Z=1)} = log \frac{K_1}{K_0} + log \frac{P(Y=1 \mid \mathbf{x})}{P(Y=0 \mid \mathbf{x})}$$

and substituting in the second term the logistic model expression, where α and β_j are the unknown parameters of the logistic model in the population we have:

$$\log \frac{P(Y=1 \mid \mathbf{x}, Z=1)}{P(Y=0 \mid \mathbf{x}, Z=1)} = \log \frac{K_1}{K_0} + \alpha + \beta_1 x_1 + ... + \beta_p x_p$$

It then follows that the logistic model for balanced data has different intercepts but equal slopes. If α^* denotes the intercept in the logistic model in the population with Z =1, it follows that:

$$\alpha = \alpha^* - \log \frac{K_1}{K_0}$$

Since K_1 is always smaller than K_0, estimation of the model on a random sample of the selected population will necessarily overstimate the default probability.

3. THE LOGISTIC MODEL FOR MATCHED CASE-CONTROL STUDIES

The logistic regression model for matched case-control studies, developed and widely used in epidemiology, may be considered as a refinement of the logistic model for balanced data. This method stratifies subjects on the basis of variables believed to be associated with the outcome. Within each stratum, samples of cases ($Y = 1$) and controls ($Y = 0$) are chosen, according to a 1 – 1 design or 1 – M design, where M number is usually no more than five.

Let K be the stratum number, n_{1k} and n_{0k} respectively the cases and the controls within the k stratum, where $k = 1, 2, \ldots, K$. The stratum-specific logistic regression model is

$$P(Y = 1 \mid \mathbf{x}, K = k) = \pi_k(\mathbf{x}) = \frac{e^{\alpha_k + \boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\alpha_k + \boldsymbol{\beta}'\mathbf{x}}}$$

where α_k represents the contribution of all constant terms within the k^{th} stratum (i.e. stratification variable or variables) and $\boldsymbol{\beta}$ the vector of the p coefficients $\beta_0 = \beta_1, \beta_2, \ldots, \beta_p$. The α_k stratum-specific parameters are nuisance terms, difficult to estimate, and therefore should be eliminated from the set of parameters on which we want to make inference. The conditional likelihood method gives consistent and asymptotically normally distributed estimates of the β_j slope coefficients. In a 1 – M match design each case is matched to M controls, so there are M + 1 units in each stratum. Letting M = 3 and denoting by x_{k1} the case and by x_{k2}, x_{k3}, x_{k4} the controls in the k^{th} stratum, the conditional probability (seen as a function of $\boldsymbol{\beta}$) that the unit with covariate x_{k1} is the case is

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k2}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k3}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k4}}}$$

4. THE LOGISTIC MODEL FOR "ROSE" DATA

Random OverSampling Examples is a new procedure developed in the R language (Menardi and Torelli, 2014), based on the generation of new artificial data according

to a smoothed bootstrap approach (Efron and Tibshirani, 1993). Let P(x) = f (x) be the probability density function on X. Let $n_j < n$ be the size of Y_j , $_j = 0,1$. A new sample is generated by the following three steps:

1. select $y = Y_j \in Y$ with probability 0.5

2. select ($\mathbf{x}_i$, y_i) in the sample such that $y = y_i$ with probability $p_i = 1 / n_j$

3. sample the vector of covariates x from the kernel probability distribution $K_{\mathbf{H}j}$ (•, $\mathbf{x}_i$) , centered on j x_i and depending on the matrix of smoothing parameters $\mathbf{H}_j$.

According to the ROSE method, in the training sample you extract a unit belonging to one of the two classes with the same probability. Then a new sample is generated in its neighborhood of width determined by $\mathbf{H}_j$ Generally, $K_{\mathbf{H}j}$ is chosen as symmetric and unimodal. Therefore the j generation of new samples for the class Y_i according to ROSE corresponds to the generation of data from the kernel density estimate of $f(x, Y_i)$, with matrix of smoothing parameters $\mathbf{H}_j$.

$$\hat{f}\left(\mathbf{x} \mid y = Y_j\right) = \sum_{i=1}^{n_j} p_i P\left(\mathbf{x} \mid \mathbf{x}_i\right) = \sum_{i=1}^{n_j} \frac{1}{n_j} P\left(\mathbf{x} \mid \mathbf{x}_i\right) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{\mathbf{H}_j}\left(\mathbf{x} - \mathbf{x}_i\right)$$

5. DATA ANALYSIS

Data are drawn from "AIDA" database, one of the most important Italian databases containing historical balance sheets as well as financial, commercial and demographic information on more than one million Italian firms. We selected all the firms in the Tuscany region having positive revenues of sales in the year 2006 and for these we extracted revenues, profits, fixed assets, financial indicators, indexes of resultant profits and current management. On May 2010, we verified their legal status: the database provides the legal status of each firm, which is periodically updated without indicating the reference date; therefore, we do not know the exact time at which a firm was declared bankrupt. The selected time interval of four years is due to the delay between the bankruptcy event and the availability of the company balance sheet. The distribution according to the legal status is shown in Table 1.

Due to the lack of information on the causes of inactivity and liquidation, we included in the analysis only the firms that are active (33,798) and bankrupt (537) . Data clearly shows the rareness of the default event (1.59%) and therefore the inadequacy of a logistic regression model due to the unbalanced data. We built a training sample, to implement the methods, and a randomly selected hold-out sample consisting of 10% of the whole sample. Since the aim of the study is to compare three different methodologies, from among the balance sheet indicators we selected those that were found to be most informative in a previous case-control study (Pierri, 2013) on the same data.

Legal status	Frequency	Percentage
Active	33798	89.23
Bankruptcy	537	1.42
Liquidation	2800	7.39
Not active	744	1.96

Table 1. Companies' distribution by state law in May 2010.

Logistic regression is estimated on three different data sets: a balanced sample with 2505 observations where the frequencies of Y_0 and Y_1 are respectively 501 and 2004; a stratified sample (2,440 observations) with strata formed on Legal Form and the first two numbers of the ATECO code (industry sector) jointly considered, where the frequencies of Y_0 and Y_1 are 488 and 1952 , respectively; and a ROSE data set of 68,000 observations where the frequencies of Y_0 and Y_1 are respectively 33,671 and 34,329. We used the ROSE routine included in R software to generate data based on the ROSE method.

	Balanced data	Balanced data2	ROSE	ROSE2	Case-Control
NP	0.00063	0.00064	2.22e-07*	-	0.00241
AC	0.12323	0.12541	0.05821	0.05813	0.10371
AC^2	-0.01069	-0.01082	-0.00636	-0.00636	-0.01176
L	-1.05589	-0.98879	-0.61510	-0.61631	-0.88630
QR	0.59219	0.60493	0.44750	0.44740	0.72559
DR	-0.00583	-0.00590	-0.00174	-0.00174	-0.00472
AT	0.47878	0.3058	0.26130	0.26927	1.02132
AT^2	-0.05038*	-	-0.00289*	-	-0.17471
EBITDAV	0.01522	0.01520	0.00747	0.00747	0.00574
AreaROC	0.7955	0.7911	0.7645	0.7648	0.7686

* p – value > 0.1

Table 2. Estimates of the coefficients and Area under the ROC curve.

The multivariable logistic model considered as explanatory variables Net Profit (NP), Asset Coverage Index (AC), Liabilities index (L), Quick Ratio (QR), Debt Ratio (DR), Asset Turnover (AT) and EBITDA value. The linearity in the odds of these variables was checked following the methodology proposed by Hosmer & Lemenshow (2013, Ch. 4): transformation of variables, applied where necessary, led to a final model including two quadratic forms (see Pierri, 2013, for details). We refer to Table 2 for a summary of our main results. The logistic model on the balanced sample indicates that for Asset Turnover the quadratic form is not statistically significant. The same holds for the ROSE sample where Net Profit is also not significant. The same table also displays the estimated coefficients for the models considering only the significant ($p < 0.05$) covariates (Balanced data2 and ROSE2).

The table reports the area under the ROC curve on the hold-out sample. The logistic model on balanced data has the greatest capability to discriminate between good and bad firms, Rose and Case-Control methods exhibit very similar results to each other.

6. CONCLUSION

Three different methodologies have been compared. On the basis of the data and the model applied, the oversampling (ROSE) and case-control studies methods seem to

give very similar results, on the other hand logistic regression on balanced data show the best predictive capabilities. We underline some particularities: ROSE allows only for continuous covariates; in case-control studies, confidence intervals are generally narrower than in standard logistic regression, but you do not have the predicted probability of bankruptcy; standard logistic regression, applied over a random balanced sample, is very easy and quick to implement. Future developments of this study will test whether stepwise model selection procedures applied to the different datasets will lead to different models.

REFERENCES

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Hosmer, D.W., Lemeshow, S., and Sturdivant R.X. (2013*). Applied Logistic Regression*. 3rd Edition. Wiley Series in Probability and Statistics.

Menardi, G., and Torelli, N. (2014). "Training and Assessing Classification Rules with Imbalanced Data". *Data Mining and Knowledge Discovery*, 28 (1), 92-122,

Pierri, F.(2013). "Valutazione del rischio di default delle PMI attraverso uno studio caso-controllo". *Gli analytics come motore per i big data, la ricerca ed il sistema paese*. Aracne Editrice, 57-68.

Pierri, F., Burchi, A., and Stanghellini, E. (2013). "La capacità predittiva degli indicatori di bilancio delle PMI". *Piccola Impresa Small Business,* Editrice Montefeltro, 85-106.

74. JOINT MODELING SCHEME BY WEIGHTING CUMULATIVE EFFECTS OVER TIME. RESEARCH ON MORTALITY FOR INSURED ELDERLY

X. Piulachs and R. Alemany
Universitat de Barcelona, Dept. of Econometrics, Riskcenter-IREA, Barcelona
xavier.piulachs@ub.edu
ralemany@ub.edu

ABSTRACT

We study longevity and usage of medical resources of a large sample of individuals aged 65 years or more who are covered by a Spanish private insurance policy. Health insurance companies accumulate big amounts of survey data on the intensity and the type of use of medical claims issued, so a historical survey can be performed on each policyholder in the form of longitudinal repeated measurements. Associated to this follow-up information, the occurrence of mortality is usually also monitored, so there is natural interest in understanding the association between both processes. Present study focuses on conducting a jointly scheme to link the weighted cumulative information of policyholder's health, which is only measured at specific time points, with time until death event.

1. INTRODUCTION

The gradual advances in medical research achieved by developed countries has lead to a larger number of years lived with disabilities, which in turn increases the demand of medical resources. Furthermore, it is well known that private insurance policyholders are generally supposed to have a higher socio-economic level compared to the rest of the population, so any inference from general mortality

tables will be probably biased. These are primary key challenges for health insurance companies, who in recent years are already facing to additional costs in order to meet the care needs in the scenario of a large cohort from age 65 onwards.

Insurers accumulate a great wealth of survey data on the intensity and the type of use of medical claims issued. Associated to the follow-up longitudinal process, the occurrence of mortality and its causes are also monitored for all policyholders. The combination of such information becomes extremely valuable to quantify the risk of medical care demand of their clients above the expected values. Building on this insight, a first survival behavior may be thought by incorporating longitudinal outcome on health status as time-dependent covariate (Andersen and Gill, 1982), but this assumption leads in practice to biased inferences (Kalbfleisch and Prentice, 2002) when considering endogenous variables, i.e., those being directly related to the event of interest mechanism and dictated by an underlying biologic response whose value is contaminated by an error term. In addition, such stochastic processes can be only observed at specific time points, so that using time-varying hazard regression models would lead to unrealistic step-profiles which remain constant between consequent measurements. In order to incorporate adequately longitudinal information into survival model, a jointly approach of both longitudinal and survival processes needs to be applied. The clue assumption on such coupling methodology is that both outcomes are linked by some common unknown random effects, assumed zero-mean normally distributed (Tsiatis *et al.*, 1995; Henderson *et al.*, 2000). Despite most of joint model frameworks focus on the strength of association between the current true value of biomarker (that is, free of error variation), and the risk of the event of interest, such relationship may not always be appropriate to combine both processes. Nonetheless, in some situations will be more informative to evaluate the risk of death at a single time point by considering the cumulative effects of subject's true health status within a period prior to that time (Sylvestre and Abrahamowicz, 2009). It is logical to conclude that this health-background should be weighted by temporal closeness with a decreasing function that decays to zero at the end of the individual's recording interval.

To the best of our knowledge, no study has evaluated how health care usage and risk of death can be modeled jointly. The main results of our analysis are: 1) to estimate

both longitudinal and survival parameters without the bias introduced by a separated analysis, 2) to confirm age and gender as most important factors affecting survival changes for a health insurance member, and 3) to find evidence of a significant association between serial weighted cumulative measurements of insurance care usage and time to event.

2. DATA AND METHODS

The motivating dataset corresponds to the information provided by a Spanish medical mutual, containing historical data which was collected between January 1st, 2006 and February 1st, 2014. In particular, our study has been limited to 39137 insurance policyholders (39.5% men and 60.5% women) who are 65 or more years when entering the study. The time origin for insured's survey is taken at the date of entry into the study, and from that moment subjects are followed within window observation. In particular, repeated measurements taken on each individual measure the annually number of selected claims which are a priori the most indicative of the event of interest, i.e., the subject's death: hospitalizations (69.5%), urgencies (27.8%) and ambulance services (2.7%). With respect to the survival time, the interest is focused on time until death. Such information is usually right-censored in our data because the majority of individuals survive beyond the end of the study. Some other insurers canceled their insurance policy and therefore they dropped out of the study automatically. The censoring process has been considered non-informative, as it was due to personal reasons not related with the event of interest.

Variable	Definition
Id	Subject identifier: i = 1, 2, . . . , 39137
Sex	Gender of the subject: 0 = Male, 1 = Female
Obstime	Observation time (years) since study entry
age0	Subject's age at study entry, where obstime = 0
log.cum	Logarithmic response of annual selected claims
time	Final observation time (years)
cens	Censoring indicator: 0 = Right-censored, 1 = Event

Table 1. Variables in the private insurer dataset (2006-2014). Variable Definition.

Table 1 presents the definition of the different variables. Due to the right skewed shape exhibited by the longitudinal outcome, a logarithmic scale has been applied to adequate for normality (Verbeke and Molenberghs, 2000).

3. A JOINT MODEL APPROACH WITH WEIGHTED CUMULATIVE EFFECTS

When conducting longitudinal studies with health indicators, many times the main issue is often not in the current value of certain biomarker, but in assessing properly its time-dependent cumulative effects within window observation. Regarding this issue, the health status forecasting of insured advanced-age population poses the challenge of modeling both the intensity of complex endogenous variable and the relevance of past values. In this work, the association between these cumulative effects and the risk of death is done by a shared random effects structure. Hence, it is assumed the dependence between two processes through an underlying latent process, so that these two responses are conditionally independent given the linking random effects.

For i-th subject, $i = 1, \ldots, n$, we define $y_i(t) = log.cumi(t) = \log\{1 + cumi\ (t)\}$ as the observed value of policyholder's annual claims. Such reading is taken at all calendar year turnovers within overall length of each policyholder's follow-up, and reflects the health status profile at a particular observation time t. At this point, it is important to highlight the fact that ultimately we want to characterize time-varying effects by considering the so-called true and unknown biomarker value, $mi(t)$, since: $y_i(t) = m_i(t) + \varepsilon_i(t)$. The true subject-specific evolutions over time are reconstructed from the observed longitudinal response by fitting a linear mixed effects model. In particular each of the n = 39137 subjects from our dataset were modeled using both fixed effects, with linear trends allowed, and random effects, with subject-specific intercept term. In addition, the observed failure time for the i-th individual is obtained as the minimum between the true survival time and the (non-informative) right-censoring time, since some policyholders may not experience death event by the end of the study. In order to introduce the mi(t) information on the risk for death event, we used a relative risk model which incorporates the weighted cumulative effects of the true historical subject profile as a time-dependent covariate. The joint model formulation is

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + b_{i0} + \beta_1 t + \varepsilon_i(t), \text{ with } b_{i0} \sim \mathcal{N}(0, \sigma^2_{b_0}) \\ h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) = \\ \qquad h_0(t) \exp\left\{\gamma_1 w_{sex} + \gamma_2 w_{age0} + \gamma_3 w_{sex \times age0} + \alpha \int_0^t \overline{w}(t-s) m_i(s) \mathrm{d}s\right\}, \end{cases}$$

where $M_i(t) = \{m_i(s), 0 \leq s \leq t\}$ is the true and complete longitudinal history of the true longitudinal outcome up to t. The measurement error terms $\varepsilon_i(t)$, i.i.d $\mathcal{N}(0, \sigma_2)$, are assumed not only mutually independent, but also independent with respect of random intercept effects b_{i0}. To avoid an underestimation in the standard errors of the parameter estimates, a B-splines baseline risk function was implemented for $h_0(t)$ in the survival submodel. In addition, the considered weight function over time $\overline{w}(.)$ is mean-zero normal distributed, with such a variance which the biomarker impact decreases as further is a health reading with respect the present moment.

Finally, the full parameter set was estimated by considering the (semiparametric) joint maximum log-likelihood from the two submodels. The main challenge is the computation task, since log-likelihood function contains an analytically intractable integral. This drawback can be overcome using EM algorithm (Dempster *et al.*, 1977).

4. RESULTS

As above commented, this study is motivated by a greater comprehension on advanced-age insured's behavior, influenced by a combination of different risk factors such as age and gender. In this regard, we have paid attention on how men's and women's health care experiences differ in the last part of their life. The joint model scheme was fitted through the jointModel function, which is implemented in the package JM (Rizopoulos, 2010) from R software. Furthermore, inferences were validated using the residuals function included in the same package. In order to assess the impact of the weight function, we considered two different models. First, the cumulative health history of recently years was considered by including the unweighted integral trajectory of true biomarker path as a liner predictor in the relative risk model, which is in practice equivalent to take $w(.) = c$. In the second

model we considered a logistic density as a weight function to handle, by recency, the time-varying health exposures. The application of joint modeling techniques by using JM package is well-documented (Rizopoulos, 2012).

JM Approach	Parameter	Value	Std. Err.	p-value
Unweighted cumulative effects	Longitudinal process			
	β_0	0.3975	0.0035	< 0.0001
	β_1	0.0317	0.0007	< 0.0001
	σ	0.6841	0.0001	
	σ_{b0}	0.3745	0.0047	
	Event process			
	γ_1	-5.3934	0.4208	< 0.0001
	γ_2	0.0814	0.0044	< 0.0001
	γ_3	0.0691	0.0053	< 0.0001
	Association			
	α	1.6058	0.0798	< 0.0001
	Fit assessment			
	Log-likelihood	-269193		
	AIC	538415		
	BIC	538544		
Logistic weighted cumulative effects	Longitudinal process			
	β_0	0.3977	0.0035	< 0.0001
	β_1	0.0319	0.0007	< 0.0001
	σ	0.6839	0.0001	
	σ_{b0}	0.3857	0.0046	
	Event process			
	γ_1	-4.9047	0.4296	< 0.0001
	γ_2	0.0909	0.0044	< 0.0001
	γ_3	0.0628	0.0053	< 0.0001
	Association			
	α	1.3064	0.0547	< 0.0001
	Fit assessment			
	Log-likelihood	-269065		
	AIC	538160		
	BIC	538288		

Table 2: Comparison on the estimates for two different joint model schemes fitted to our dataset of private health insurance company.

The resulting maximum likelihood estimates of main parameters from both submodels are presented in Table 2. Besides, as an essential result to be meaningful the linkage between longitudinal and event outcomes, the parameter α is shown strongly significant and with positive sign. Therefore, an increasing relationship between frequency of use of considered insurer services and risk of death can be inferred. In spite of two models behave similar results, the joint approach involving weighted cumulative effects reaches greater likelihood and also less penalization by Akaike and Bayesian criteria.

Furthermore, this second propose presents an association estimate with a considerably smaller standard error than the unweighted model. The weighted-exposure model displays that a unit increase in the area under the log-cumulated effects correspond to a 1.3-fold increase in the risk for death event, holding the other covariates constant.

5. CONCLUSIONS

The main results of our analysis are: 1) we confirm age and gender are main factors influencing changes in survival for a health insurance member, and 2) we find evidence of a significant association between serial measurements of cumulative private insurance care usage and longevity. In summary, our contribution shows that an increase in health care usage intensity is negatively associated with survival, but that its influence varies as usage accumulates and depending on other factors such as sex and age, as well as previous insurance conditions.

REFERENCES

Andersen, P., and Gill, R. (1982). "Cox's Regression Model for Counting Processes: A Large Sample Study". *Annals of Statistics*, 10, 1100-1120.

Dao, H., Godbout, L., and Fortin, P. (2014). "On the Importance of Taking End-of-Life Expenditures into Account when Projecting Health-Care Spending". *Canadian Public Policy*, 40, 45-56.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Denuit, M., and Frostig, E. (2008). "First-Order Mortality Basis for Life Annuities". *The Geneva Risk and Insurance Review*, 33, 75-89.

Kalbfleisch, J., and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition. New Jersey: John Wiley and Sons.

Rizopoulos, R. (2010). JM: "An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data". *Journal of Statistical Software*, 35, 1-33.

Rizopoulos, R. (2012). *Joint Models for Longitudinal and Time-to-Event Data. With Applications in R*. Boca Raton (Florida): Chapman and Hall/CRC Biostatistic Series.

Sylvestre, M.-P., and Abrahamowicz, M. (2009). "Flexible Modelling of the Cumulative Effects of Time-Dependent Exposures on the Hazard". *Statistics in Medicine*, 28, 3437-3453.

Tsiatis, A., DeGruttola, V., and Wulfsohn, M. (1995). "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS". *Journal of American Statistical Association*, 90, 27-37.

Verbeke, G., and Mohlenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

75. REAL OPTIONS VALUATION UNDER UNCERTAINTY

M. Lambert and F. Platania
University of Liège

M. Moreno
University of Castilla-La Mancha

ABSTRACT

In this paper we develop a novel valuation model and methodology to value a pharmaceutical R&D project based on real options approach. The real options approach enables the possibility of optimally abandon the project before completion whenever the investment cost turns out to be larger than the expected net cash flow stream. On the other hand, the proposed model accounts for two different sources of uncertainty, those are technical and economic risk. This model incorporates a novel economic state vector where each economic state captures the interaction among different market and economic forces using Fourier series as the particular basis for the economic function space. In this sense, Fourier series are considered as an aggregate of forces playing a relevant role in the process evolution determining the cash flow structure and also allowing us to properly define an economic scenario where the project will be developed.

1. INTRODUCTION

The pharmaceutical industry is one of the most dynamic and research-intensive industries in the world. One distinctive characteristic of this sector is the high level of investment in research and development, in fact the pharmaceutical industry has one of the highest R&D budget to sales revenue ratio across industries. The aim of this paper is to provide a comprehensive tool and methodology to value an R&D

project considering that it is subjected to technical and economic uncertainty. Here we focus the attention on pharmaceutical R&D projects, however the model and methodology used can be easily extrapolated to any industry, say for instance mining projects. The pharmaceutical market is extremely complex and has divided the public opinion in several controversial topics such as animal testing, drug prices, lack of research interest for certain diseases, public funding, and so on. As any other private company, pharmaceutical companies are ultimately focused on increasing shareholder value. The public perception that privately research funding is solely motivated by profit has increased the friction between shareholders' return expectations and the public notion of fairness.

On this regard, it is important to point out that no matter how big a pharmaceutical company is, it can only cover a small portion of breakthrough R&D projects. Therefore, most pharmaceutical companies have to choose with financial wisdom each project because they simple cannot afford to invest when the affected population is too poor to buy the drug or the market niche is just too small to achieve a reasonable return on the investment. Recently, the 2014 Ebola outbreak has revealed the lack of resources and effort assigned to fight this virus while it was limited or contained within the African border, and the increased interest when the virus crossed the European and American border and "opened a new market". In this paper we are not going to discuss such controversial ethical issues, however our proposed model incorporates a novel economic state vector where each economic state captures the interaction among different market and economic forces using Fourier series as the particular basis for the economic function space. Hence, our model can be used to depict any extreme economic situation and properly value an R&D project targeting such market. Furthermore, since most drugs introduced by the pharmaceutical industry are developed with some contribution from the public sector, see for instance Cockburn and Henderson (2000), our model can be used to determine the appropriate amount of taxpayer's money to allocate in a specific project. On this regard, it is worth to mention that public opinion is a strong force which can heavily affect the project value, its effects can also be modelled with the appropriate terms in the Fourier expansion.

Our approach is close to the work of Berk *et al*. (2003) and Schwartz (2004) where the cash flows from the R&D project are modelled. In more detail, Berk *et al*. (2003) develops and analyses a single R&D investment project modelling the cash flows

from the project with two stochastic processes, one of them tracking any possible catastrophic event and the other process modelling the conditional cash flows the project would have produced if it were completed. Schwartz (2004) implements a simulation approach to value patents and patent-protected R&D projects assuming two stochastic differential processes, one of them for the cost-of-completion and the other for the cash flows generated from the project, and introduces the probability of any catastrophic event with a Poisson probability. In this paper we consider the net cash flow as the underlying variable, however since this variable takes into consideration the production and marketing cost it could yield a negative cash flow stream. Therefore, we assume that the net cash flow of a successful project is given by an Arithmetic Brownian motion process plus a time dependent component depicts by the Fourier series. The objective of this paper is to provide a powerful and flexible valuation model and technology accounting for technical and economic risk and considering all those relevant forces playing a significant role in the project valuation and decision making process.

2. R&D VALUATION MODEL

Consider, for instance, a pharmaceutical R&D project for the development of a new drug. The very nature of such project and the potential impact on human health make the pharmaceutical industry quite unique and risky. The overall project's life can be divided into two mayor phases, firstly the research and development phase and secondly the market phase. During the early-stage of the research and development phase, a new compound which may potentially derive into a marketable drug is either discovered or designed. Once the compound is successfully identified as a potential drug and synthesized the project moves to the next stage. During the preclinical and clinical development the drug must successfully complete a number of well regulated stages. Firstly, the preclinical stage covers the laboratory and animal testing, and it is normally during this stage when the company applies for a patent. If and only if the drug successfully completes the preclinical stage, it accesses the clinical stage which can be divided into clinical phase I, II, and III. During the clinical phase I, the drug or treatment is tested in a small group of healthy volunteers in order to determine the safe dosage, evaluate its safety, and to identify possible side effects and toxicity. During the clinical phase II, the drug or treatment is tested on a

relative large group of subjects (100-300) with the condition that the drug is intended to treat in order to further evaluate its safety and efficacy. Finally, the clinical testing phase III consists of large scale trials, usually a few thousands, to confirm the safety and efficacy of the drug or treatment and to further monitor possible side effects. The final stage in the research and development phase is the marketing approval, once again if and only if the drug has successfully completed each and every preceding stage, the regulatory authority decides whether the drug is approved for patient use or not. If the marketing approval is granted the project moves to the market phase where the appropriate marketing strategy should be established and the product is market launched.

During the patent's life, the company is entitle to a set of exclusive rights protecting the project from market competitors for a limited period of time. However, market competition is not the only force that jeopardises the successful completion of the project. It is well established in the literature, see for instance Brealey and Myers (2000), that an R&D project faces two different sources of risk, those are the economic and technical risk. Technical or technological risk takes into account the inherent uncertainty about the successful completion of each stage during the drug development phase, for instance, an extreme side effect during the clinical testing would lead to a failure yielding nothing but a worthless project. On the other hand, economic risk deals with both market uncertainty such as sales volume, pricing levels, market competitors, and other economic factors such as interest rates, inflation, growth rate.

2.1. Technical uncertainty

Technical or technological risk is the primary source of uncertainty during the drug development process, in fact, most drugs undergoing the preclinical and clinical stage do not obtain the regulatory authority's approval. Since each stage must be preceded by the successful completion of the previous one, the failure of one stage produces the overall project termination. However, once the drug successfully passes the preclinical and clinical test and finally achieves the regulatory authority's approval technical risk vanishes. Considering that the outcome of a failure is a worthless project, it is widely spread the use of a Poisson process to model technical or technological risk (see for instance Pennings and Sereno (2011), Schwartz (2004),

among others). The Poisson probability mass function (pmf) is given by $f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$, where $\lambda > 0$ is the Poisson parameter, and $k = 0, 1, 2, \ldots, \infty$ defines the number of events. Since a failure event can only occurs once, that is after failing one stage the project is abandoned, the pmf is determined by the mutually excluding events $k = 0, 1$, where $f(k = 0; \lambda)$ is the probability that the project successfully passes the stage under consideration, and $f(k = 1; \lambda)$ is the probability of failure, hence $1 = f(k = 0; \lambda) + f(\mathrm{k} = 1; \lambda) = \mathrm{e}^{-\lambda} + \lambda \mathrm{e}^{-\lambda}$. Assuming that failing a stage sinks the project value to zero, technical risk can be consider as a premium over the risk free rate, therefore, during the development process the discount factor is given by $e^{-rdt} = e^{-(r+\lambda)t}$, where λ represents the annual rate of failure and r the risk-free rate. Note that, as stated above, technical risk vanishes after the regulatory authority's approval, hence this risk premium is only valid during the drug development phase.

2.2. Economic and market uncertainty

So far things are fairly easy but we have only dealt with technical risk. Economic risk takes into account those factors affecting market conditions, not determining the successful completion in technological terms but defining the cash flow structure of a successful project which gives rise to the project's abandon option. In this fashion, economic risk not only comprehends macro and microeconomic figures but also certain project specific forces and circumstances driving the cash in and out flow, for instance an outbreak of influenza would drive an increase in market sales for those specific medicines or the 2014 Ebola outbreak that pushed the use of the experimental drug "ZMapp" in humans. Note that we have intentionally used the sentence "successful project", that is because we have divided the project into two mayor phases, the research and development and the market phase. As stated above, during the research and development phase technical risk is the dominant source of uncertainty and it vanishes as the drug successfully overcomes every single stage in the development process and finally achieves the corresponding approval, those projects reaching the market phase are the "successful projects". Once the drug reaches the market phase, there are several forces playing a significant role and we have called this source of uncertainty economic risk, however it is important to remember that only a "successful project" will face economic risk.

We can easily realise that measuring economic risk is not a trivial endeavour, in fact, creating a framework where every force affecting the project is considered is literally impossible. On this regard, it is common use to model the evolution of the project or the evolution of the cash flow as a stochastic differential equation given as: $dC_t = \mu(C, t)dt + \sigma(C, t)dW$, where the process can take the form of a Geometric Brownian motion, Arithmetic Brownian motion, or an Ornstein-Uhlenbeck process. We can also find a more realistic and sophisticated framework, as the one proposed by Schwartz (2004), where the author models both the cash flow and the cost of completion with a stochastic differential equation.

As long as we stick with one stochastic factor, all these models share that the only source of uncertainty comes from a random walk weighted by $\sigma(C, t)$, that is the diffusion term. It seems fairly obvious that a simple diffusion model cannot account for a realistic variety of forces playing a key role during the project's market phase. In particular, neither of these models can properly account for any seasonal component which, for instance, plays a primary role in the outbreaks of the flu, plus neither consider the effect of the business cycle nor any other relevant force. At this point it is worth to wonder whether such models are an oversimplification and which forces do really make an impact in terms of project valuation. Of course there is not one right answer, each project must be analysed in excruciating detail to determine the appropriate set of relevant forces, but it seems fair to conclude that a simple diffusion model is just a naive simplification of the market structure.

In what follows we consider that the net cash flow stream, C_t, of a successful project is given by a latent variable, Y_t, depicted by an Arithmetic Brownian motion process plus a time dependent component described by the Fourier series, that is: $C_t = f(t) + Y_t$; $dY_t = \mu dt + \sigma dW_t$; $f(t) = P\infty - \infty\, Re\,[Ane^{inwt}]$, where $\{A_n \in \mathbb{C}\ \forall n\}$, $\{(\mu, \sigma, \omega) \in \mathbb{R}\}$, and we only consider the real part in the Fourier expansion since it is the part that makes economic sense. Note that, applying Ito's lemma to C_t, the net cash flow dynamic is depicted by $dC_t = \left(\mu + \frac{df}{dt}(t)\right) dt + \sigma dW_t$. The net cash flow stream takes into consideration the production and marketing cost, in consequence it could yield a negative rate, thus an Arithmetic Brownian motion process is a suitable representation of the latent underlying process. Under the risk neutral probability $\mathbb{P}^Q$, the solution of the latent underlying process is given as: $Y_t = Y_0 e^{rt} + \sigma \int_0^t e^{r(t-s)} dW_s^Q$, and W_t^Q the net cash flow at any given time t is represented by $C_t = Y_0 e^{rt} + f(t)$

$+ \sigma \int_0^t e^{r(t-s)} dW_s^Q$, where W_t^Q is a standard Wiener process under the risk-neutral measure $\mathbb{P}^Q$.

The cash flow stream starts when the R&D project is market launched, before this stage the process describes the net cash flow that the project would have produced if it were successfully completed. Once the medicine or treatment is market launched the value of the project depends exclusively on the net cash flow generated. Hence, using the Merton (1973) no-arbitrage technique the project value, $V(C_t, t)$, must satisfy the following partial differential $\frac{\partial V}{\partial t} + rC\frac{\partial V}{\partial C} + \frac{\sigma^2}{2}\frac{\partial^2 V}{\partial C^2} - rV = 0$, subject to the appropriate terminal condition $V(C, T)$, where T represents the patent expiration.

The novel component in this model is the ad hoc incorporation of the Fourier series, $f(t)$, accounting for any economic, market, and specific force affecting the project and not captured by the underlying stochastic differential equation. In this sense the Fourier series should be considered as an aggregate of forces playing a relevant role in the process evolution and determining the cash flow structure. Therefore, Fourier series allows us to properly define a scenario where the project will be developed. Of course, such scenario is tailor made based on the characteristics of each project, the influence and exposure to certain forces, and so on. Note that, so far we just have introduced a fully deterministic ad hoc scenario, we can define it as much sophisticated as we want increasing the number of forces affecting the evolution of the project. In fact, Fourier series provides a great deal of flexibility as, by Carleson's theorem, it converges almost everywhere for a L^2 function, but still a deterministic process. So, in terms of risk management is fair to wonder whether a deterministic scenario is enough or an adequate instrument to actually addresses economic risk. In order to tackle this situation, let us represent the economic uncertainty by the state vector $\Phi^{(j)}$ with $j \in \mathbb{N}$, where each state defines a case scenario depicted by a concrete selection of terms in the Fourier expansion and represents the aggregate of forces. It is important to stress out that a state scenario does not attempt to replicate a precise future outcome but rather establishes an alternative future development, in this sense the state vector Φ represents a set of "alternative worlds". Each state determines the cash flow structure of a successful project and consequently the managerial decision of ceasing or continuing the project. Thus, the expected patent value conditional to a certain economic state is given as $\mathrm{V}(\mathrm{t}, C_t, \mathrm{I}_t; \Phi^{(j)} = \mathrm{E}[\mathrm{V} | \Phi^{(j)}]$ with

$j \in \mathbb{N}$, where C_t and I_t represent the net cash flow structure once the drug obtains the marketing approval and the investment structure during the research and development phase, respectively.

Note that the conditional patent value is constrained to the future development of certain state, which of course is uncertain. Therefore, since Φ is defined as a discrete state vector, an essential piece of the puzzle is the appropriate definition of its mass probability function. On this regard, Huchzermeier and Loch (2001) define a one-dimensional parameter i to model the product performance. The authors claim that this performance may unexpectedly improve with probability p, or it may deteriorate with probability $(1-p)$ and they generalize the binomial distribution by allowing the performance "improvement" and "deterioration" over N performance states. We can easily accommodate a similar probability mass function defining two states in the economic state vector, say for instance, "Optimistic" and "Pessimistic" states. However, as stated above, each state represents the aggregate of forces acting over the project, and therefore is very project specific. Note that each state can be defined in several ways, we can tailor made it based on our own expectations, we can define it based on analyst expectations, and so on. Hence, let us define the state vector probability mass function in general terms as $g\,(\Phi^{(j)}) = Pr\,(\Phi = \Phi^{(j)}) = p_j$, where p_j represents the probability that the state $\Phi^{(j)}$ turns out real. Hence, under this framework the patent value is determined as: Patent Value $= \frac{\sum_j V\left(t, C_t, I_t; \Phi^{(j)}\right) \cdot g\left(\Phi^{(j)}\right)}{\sum_j g\left(\Phi^{(j)}\right)}$.

3. CONCLUDING REMARKS

In this paper we have developed a novel valuation model and methodology to value a pharmaceutical R&D project based on real options approach. Firstly, considering that the outcome of a failure is a worthless project, we have modelled technical or technological risk using a Poisson process. Therefore, during the research and development phase there is a premium of λ over the risk free rate which is related to this sort of risk.

On the other hand, during the market phase technical risk vanishes and it is the economic risk the dominant source of uncertainty. In order to address economic risk we have incorporated a novel economic state vector where each economic state

captures the interaction among different market and economic forces using Fourier series as the particular basis for the economic function space. Fourier series allows us to properly define an economic scenario where the project will be developed and it is considered as an aggregate of forces playing a relevant role in the process evolution and determining the cash flow structure. On this regards, Fourier series is a powerful mathematical instrument which allows us to define the economic state scenario as much sophisticated as we want increasing the number of forces affecting the evolution of the project.

The model and methodology presented in this paper constitute a powerful and yet simple valuation instrument with strong practical applications. There are several forces, both economic and technical, driving the drug development process that are not fully understood. Our proposed model tackles all those forces playing a significant role in the project valuation process in a very simple manner and provides a comprehensive tool for the decision making process. The model and methodology here proposed can be easily extrapolated to any other industry or corporate project.

REFERENCES

Berk, J.B., Green, R.C., and Naik, V. (2004). "The Valuation and Return Dynamics of New Ventures". *Review of Financial Studies*, 17, 1-35.

Brealey, R., and Myers, S. (2000). *Principles of Corporate Finance*. McGraw-Hill.

Cockburn, I., and Henderson, R. (2000). "Publicly Funded Science and the Productivity of the Pharmaceutical Industry". *National Bureau of Economic Research*, 1-34.

Huchzermeier, A., and Loch, C. (2001). "Project Management Under Risk: Using the Real Options Approach to Evaluate Flexibility in R&D". *Management Science*, 47 (1): 85-101.

Merton, R.C. (1973). "Theory of Rational Option Pricing". *Bell Journal of Economics and Management Science*, 4, 1, 141-183.

Pennings, E., and Sereno, L. (2011). "Evaluating Pharmaceutical R&D under Technical and Economic Uncertainty". *European Journal of Operational Research*, volume 212, issue 2, 16 July 2011, 374-385.

Schwartz, E.S. (2004). "Patents and R&D as Real Options". *Economic Notes*, 33: 1, 23-54.

76. BEHAVIORAL EQUILIBRIUM EXCHANGE RATES: RE-ESTIMATION FOR THE EURO AREA COUNTRIES

L. Issever Grochová
Mendel University in Brno, Department of Economics, Czech Republic
ladislava.grochova@mendelu.cz

K. Plecitá
Mendel University in Brno, Department of Finance, Czech Republic
klara.plecita@mendelu.cz

ABSTRACT

This paper focuses on estimating the long-run determinants of the real effective exchange rates for the euro area countries towards the rest of the euro area over the period 1999-2014. Our approach is based on the Behavioral Equilibrium Exchange Rate (BEER) concept, which uses cointegration to determine the long-run fundamentals of real exchange rates. Terms of trade, relative productivity differential between tradable and non-tradable sector and net foreigner assets are widely used as the main factors affecting the intra-euro-area real effective exchange rates. In contrast to the main body of empirical literature our estimates indicate a negative, albeit weak, relationship between net foreigner assets and intra-euro-area real effective exchange rates. Possible explanation for this phenomenon might be the accumulation of economic imbalances across the euro area during the period analyzed.

1. INTRODUCTION

The concept of the Behavioral Equilibrium Exchange Rate (BEER) was developed as a reaction to the inability of the Purchasing Power Parity (PPP) to explain the

observed real exchange rate behavior. In contrast to the PPP theory, the BEER approach uses cointegration to estimate the equilibrium real exchange rate. Even if there is no consensus concerning the long-run determinants driving the equilibrium exchange rates, the vast majority of existing empirical literature, including our previous estimates on BEERs inside the euro area over 1999-2011 published in Grochová and Plecitá (2013), identify terms of trade, productivity and net foreigner assets as the fundamental determinants of real effective exchange rates. However, estimation results vary significantly over countries and time spans. Since it can be expected that the resolution of the accumulated euro-area economic imbalances can directly affect both actual and equilibrium real exchange rates inside the euro area, in this paper we reestimate the BEER for the panel of the euro area countries over the period 1999-2014 to verify the stability of our previous results.

2. DETERMINANTS OF THE BEHAVIORAL EQUILIBRIUM EXCHANGE RATE

As previously mentioned the principal determinants of the BEER are productivity, terms of trade and net foreigner assets. The Balassa-Samuelson hypothesis is routinely used as an explanation for exchange rate deviations from the PPP. According to the Balassa-Samuelson hypothesis, growth of productivity in the tradable sector leads, ceteris paribus, to wage increases which in turn lead to higher prices of nontradables and real exchange rate appreciation. The empirical evidence for the positive impact of tradable productivity growth on real exchange rate is at least mixed (for overview see Peltonen and Sager, 2009). Authors who have identified strong negative relationship between tradable productivity and exchange rate explain it for example on the basis of productivity increasing skill-biased technological change, which lowers demand for low-skilled labor and therefor the overall price level (Gubler and Sax, 2011) or on the basis of low level of labor market efficiency (Sheng and Xu, 2011).

Another important long-term determinant of the equilibrium exchange rate are terms of trade which are expected to have a positive effect on real exchange rates. An increase in relative prices of domestic exports is followed by a substitution effect-domestic producers shift their focus to the traded sector, which triggers first wage increases in the tradable sector and later (due to the wage equalization process)

in the non-tradable sector. This increases the overall domestic price level and appreciates the real exchange rate. Nevertheless, the substitution effect can be overridden by a wealth effect, which causes real exchange rate depreciation. The empirical evidence on the impact of terms of trade on the real exchange rate is ambiguous (see Frait *et al.*, 2007 for an overview).

Traditionally it is expected that improvement in net foreigner assets increases household wealth, which enables higher absorption and current account deficits. The effect of improving (raising) net foreigner assets on real exchange rate is positive, either by increasing the overall domestic price level or by an appreciation of the nominal exchange rate (Borgy *et al.*, 2008).

3. DATA AND METHODOLOGY

Our panel consists of 12 euro area countries (EA12), which adopted the common currency euro (in cash) in 2002: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal and Spain. Since not all data are individually available for Belgium and Luxembourg, in our estimation we use joint data for the Belgium-Luxembourg Economic Union (BLEU).

The real effective exchange rate *(reer)* vis-à-vis the rest of the EA12 is calculated as a weighted geometric average according to the methodology described by Schmitz *et al.* (2012) and is quoted in indirect form; hence an increase (decrease) in *reer* represents an appreciation (depreciation). The quarterly data on nominal effective exchange rates and harmonized indices of consumer prices used for *reer* computation are obtained from the European Commission and Eurostat, respectively. Weights used by the European Commission to calculate the nominal effective exchange rates are also used to calculate relative terms of trade *(tot)*. Relative terms of trade are defined as the domestic terms of trade, i.e. ratio of domestic export prices and domestic import prices, divided by the geometrically averaged terms of trade for the rest of the EA12. Quarterly data on export and import indices are both taken from Eurostat. The relative productivity differential *(prod)* captures the productivity of tradables over non-tradables for a given country relative to the average productivity of tradables over non-tradables for the rest of the EA-12. Value added per employment

in agriculture, industry and manufacturing is used to compute productivity in the tradable sector, whereas value added per employment in services, building and construction represents the non-tradable sector. Since the sectorial value added and employment data from AMECO are available only on the yearly basis, cubic spline interpolation has been used to obtain quarterly data. This technique has been also applied to net foreigner assets data, since Eurostat publishes only yearly data prior the year 2004. In our model we use the ratio of net foreigner assets to GDP *(nfa)*. Following the goal we estimate the long-run ln*reer* (aka BEER) as a function of ln*tot*, *nfa* and ln*prod*:

$$\ln reer_{it} = \mu + \beta_0 \ln tot_{it} + \beta_1 nfa_{it} + \beta_2 \ln prod_{it} + \varepsilon_{it}$$

Preliminary to the estimation process, all variables are tested for stationarity - to make the results more robust we use the test developed by Im *et al.* (2003), as well as the complementary test with reversed null and alternative hypotheses by Hadri (2000). The correct number of lags is chosen with the Bayesian information criterion. We find that all variables have one unit root, while error term is an I(0) process which suggests a possible long-run relationship.

Table 1 shows the results for four alternative panel data estimation procedures-pooled OLS, fixed effect model (FEM), random effect model (REM) and feasible GLS (FGLS), and diagnostic tests. In order to choose the correct model specification we perform panel diagnostic tests (see Table 1).

4. EMPIRICAL RESULTS AND DISCUSSION

Our estimation results for the panel of the EA12 countries (see Table 1) indicate a statistically significant cointegrated relationship between real effective exchange rate and its long-term determinants-relative terms of trade, relative productivity differential and net foreigner affects; however the sign and magnitude of our point estimates differ from our previous results published in Grochová and Plecitá (2013). Previously we have identified the *tot* elasticity to be of double magnitude and the *nfa* to have a positive impact on *reer*.

	Pooled OLS	FEM	REM	FGLS
ln*tot*	0.2080 (0.0380)***	0.1545 (0.0362)***	0.1549 (0.0361)***	0.2420 (0.0262)***
ln*prod*	0.0297 (0.0052)***	-0.0460 (0.0146)***	-0.0372 (0.0135)***	-0.0117 (0.0037)**
nfa	-0.0004 (0.0000)***	-0.0002 (0.0000)***	-0.0002 (0.0000)***	-0.0001 (0.0000)***
μ	4.6140 (0.0015)***	4.6188 (0.0012)***	4.6187 (0.0099)***	4.626 (0.0009)***
N	693	693	693	693
R-sq	0.5488	0.3280	0.3589	0.3957
BIC	-2636.5	-3317.2	-5768.3	-2899.4
test	statistics		p-value	
F-test	113.42		0.000	
B-P LM test	6897.63		0.000	
Hausman test	5.61		0.132	
Wald test	1016.76		0.000	

Note: Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001.

Table 1. Estimation of BEER.

The newly estimated positive elasticity of relative terms of trade is positive, which means that an improvement (increase) in the *tot* leads to an appreciation of the real effective exchange rate. Thus indicating that in the case of the euro area, the income effect dominates the substitution effect. Even if our new results indicate, that there is a long-run equilibrium relationship between relative productivity differential and intra-euro-area real effective exchange rates, the sign of this relation is negative which contradicts the Balassa-Samuelson hypothesis. Likewise result is presented in Grochová and Plecitá (2013), however in this previous paper we used a different construction of the relative productivity differential defined as an effective ratio of the non-tradable sector productivity to the tradable sector productivity. Thus the point estimates of the relative productivity differentials in these two studies are bound to have opposite signs and their size cannot be meaningfully compared. Negative relationship between the productivity differential and the real exchange rate is identified among others by Peltonen and Sager (2009) and Gubler and Sax (2011). These authors mention as a possible explanation for their findings pricing to market strategies and skill-biased technological change,

respectively. In our opinion both of these explanations seem to be reasonable for the euro area countries.

The fact that we find this time a robust negative relationship between net foreigner assets and intra-euro-area real effective exchange rate is puzzling, even though it appears to be relatively small. Economic theory expects a positive relationship between net foreigner assets and real exchange rate and this expectation is often affirmed by empirical analysis. However, our paper is not the only one to report this negative relationship for the euro-area countries. Borgy *et al*. (2008) have investigated how a membership in the euro area affects exchange rate determination. Their main findings are that equilibrium real exchange rates among the group of the euro-area countries and the group of the not-euro-area OECD members are determined by relative productivity and net foreigner assets. The effects of these variables on the real exchange rate are estimated to be positive with the exception of the net foreigner assets in the euro area, which exhibit a relatively small and namely negative effect on *reer*. However, Borgy *et al*. (2008) have not come up with a plausible explanation for the negative relationship. In our opinion the negative relationship between *nfa* and *reer* identified in our estimation might mirror the accumulation of economic imbalances across the euro area members. The logic of this argument is as follows: around the creation of the euro area the catching-up (peripheral) countries were expected to have higher growth potential than the core countries, therefore the catching-up countries were more appealing to foreigner capital. Moreover, the decrease of interest rates in the catching-up countries and the increase of external credit from the core countries with higher savings helped to increase their aggregate demand (mainly consumption) and fuel speculative bubbles, which raised the domestic price level and wages in both the non-traded and traded sector. This led in the catching-up countries to the appreciation of their real effective exchange rates vis-à-vis the rest of the EA12, to the loss of their price competitiveness in the tradable sector and to persistent current account deficits. Since our panel covers only 16 years, it is possible, that the estimated cointegration relationship between *nfa* and *reer* does not reflect the long-run economic equilibrium, but is merely describing accumulation of external imbalances inside the euro area. Therefore, we recommend applying more methods of equilibrium exchange rate estimation simultaneously (e.g. fundamental equilibrium exchange rate or natural equilibrium exchange rate) to obtain more robust results.

5. CONCLUSION

We have identified terms of trade, relative productivity differentials and net foreigner assets as the main determinants driving the long-run equilibrium effective exchange rate of the euro area countries vis-à-vis the rest of the euro area. Our findings are with line with our previous research (Grochová and Plecitá, 2013) with the exception of the impact of the net foreigner assets on intra-euro-area real effective exchange rates, which is estimated to be negative confirming inconsistency in results of many previous studies.

REFERENCES

Borgy, V., Lahreche-Revil, A., and López-Villavicencio, A. (2008). "Exchange Rate Dynamics in Monetary Union: An Empirical Assessment". *Ten Years of a Common Currency: Achievements and Policy Challenges in the Euro Area*.

Breusch, T.S., and Pagan, A.R. (1980). "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics". *Review of Economic Studies* 47, 239-253.

Grochová. L., and Plecitá, K. (2013). "On Exchange Rate Misalignments in the Eurozone's Peripheral Countries". *Numerical Analysis and Applied Mathematics ICNAAM 2013: International Conference of Numerical Analysis and Applied Mathematics.*

Gubler, M., and Sax, C. (2011). "The Balassa-Samuelson Effect Reversed: New Evidence from OECD Countries". *WWZ Discussion Paper*, 9.

Hadri, K. (2000). "Testing for Stationarity in Heterogeneous Panel Data". *Econometrics Journal* 3, 148-161.

Im, K.S., Pesaran, M.H., and Shin. Y. (2003). "Testing for Unit Roots in Heterogeneous Panels". *Journal of Econometrics* 115, 53-74.

Kutasi, G. (2013). "The Reverse Balassa–Samuelson Effect in the Euro Zone". *Közgazdaság*, 8 (1), 167-173.

Frait, J., Komárek, L., and Melecký, M. (2004). "The Real Exchange Rate Misalignment in the Five Central European Countries". *Warwick Economic Research Papers*, 739.

Peltonen, T.A., and Sager, M. (2009). "Productivity Shocks and Real Exchange Rates: A Reappraisal". *European Central Bank Working Paper Series*, 1046.

Schmitz, M., De Clercq, M., Fidora, M., Lauro, B., and Pinheiro, C., (2012). "Revisiting the Effective Exchange Rates of the Euro". *European Central Bank Occasional Paper Series*, No. 134.

Sheng, Y., and Xu, X. (2011). "Real Exchange Rate, Productivity and Labor Market Rigidities". *Journal of International Money and Finance* 30, 587-603.

77. SOLVING THE INTEGRATED LOCATION ROUTING PROBLEM CONSIDERING UNCERTAINTY AND RISK FACTORS

C.L. Quintero and Á.A. Juan
IN3 - Open University of Catalonia, Computer Science Department, Spain
{cquinteroa, ajuanp}@uoc.edu

ABSTRACT

The Location Routing Problem (LRP) involves three key decision levels (strategic, tactical, and operational) in logistics systems. In this paper a simulation-based approach is proposed to deal with the LRP, under uncertainty conditions, in an integrated way. This method uses a random choice of the depot and a fast routing heuristic to determine the depot location to be opened. Then, a randomized routing heuristic considering safety stocks to minimize risk of route failures is used to solve the corresponding Capacitated Vehicle Routing Problem with Stochastic Demand (CVRPSD). Finally, a numerical example including a risk-based comparison of solutions is presented to show the validity of the method.

1. INTRODUCTION

The Location Routing Problem (LRP) is the combination of the Facility Location-Allocation Problem (FLAP) and the Vehicle Routing Problem (VRP). It implies all decision levels in logistics systems: strategic and tactical levels correspond to FLAP while operational decisions are mostly related to the VRP. Since the FLAP and the VRP are known to be NP-Hard, the LRP is also NP-Hard —which implies that it is not possible to find optimal solutions for large instances in a reasonable computational time. LRP can be used in supply chain design (Cordeau *et al.* 2006) as well as in urban logistics design (Muñoz-Villamizar *et al.* 2013). In most of the works available in the literature, the FLAP is firtsly solved and then the solution is used as an input

for solving the associated VRP, this has been made by means of mixed integer linear programming, heuristic methods, or a combination of them. Being sequential approaches, they do not consider the integrated problem. Therefore, solutions tend to be sub-optimal.

In this work, a simulation-based heuristic, which combines random selection of the node with a fast routing heuristic, is used to solve the FLAP under uncertainty conditions (stochastic demands). Notice that routing costs are also involved in this decision. Then, a more efficient heuristic is used to solve the routing phase using risk-aware safety stocks to deal with stochastic demands.

The remaining of this paper is organized as follows: in Section 2 a brief literature review is provided; Section 3 includes an overview description of the proposed method; Section 4 discusses some numerical experiments to test the validity of the method; finally, Section 5 presents the conclusions and perspectives for future work.

2. LITERATURE REVIEW

Most of the works available in the literature regarding supply chain design correspond to deterministic cases. In Muñoz-Villamizar *et al.* (2013), a randomized heuristic is used to solve the location of urban distribution centers with deterministic demands using real data from Saint Etienne in France. The location phase is solved using the center of mass of customers and then the associated VRP is tackled by a simulation-based heuristic. More recently stochastic parameters have been considered. In Seyedhosseini *et al.* (2014) random disruptions in facilities are considered in a three-level supply chain, the problem was solved by means of a genetic algorithm. In Ahmadi-Javid and Seddighi (2013) a LRP with production and distribution disruption risk for a two-echelon supply chain is solved by means of a mixed-integer programming model for small instances and by simulated annealing for mid-large size instances. Herazo-Padilla *et al.* (2013) proposed a randomized heuristic to solve depot location, Ant Colony Optimization (ACO to solve the routing phase and Simulation in Arena to calculate the expected costs of the routes. In Juan *et al.* (2011) safety stocks are introduced as a way to deal with uncertain demand in vehicle

routing problems. Reliability (risk analysis) of solutions and expected cost are calculated using Monte Carlo Simulation. For a more detailed literature review on location-routing problems, the reader is referred to Prodhon and Prins (2014).

3. SOLUTION METHOD

To deal with the Location issue of the problem, *k* depots are randomly selected from all the possible locations (*n*). For each depot, the associated VRP is solved by means of the Clarke & Wright Savings heuristic (CWS). This heuristic generates a set of *k* feasible solutions for the LRP with deterministic demands. Each solution is placed in a priority list according to its costs. To tackle the routing phase, the depot placed at the top of the list is selected and the associated CVRPSD is solved by using a simulation-based heuristic with the following considerations:

1. Stochastic demand follows a lognormal distribution. Since real life demands are mostly associated with non-negative values, therefore they should be modeleled by using a distribution offering non-negative values or asymmetries due to long right-hand tails (log-normal, weibull, etc.). The parameters of the distribution will be obtained using $\mu = E[D_i]$ *and* $Var[D_i] = k\, d_i$, with $0<k<1$.

2. The capacity (*W*) of each truck is reduced in a given percentage (*p*)
   This reduction will be used as a safety stock to deal with the uncertainty of the demand, thus minimizing the risk of suffering a rotue failure. In this case $p = 0.1$

3. Solve the associated CVRP with reduced capacity *W'* by means of the biased-randomized version of CWS proposed in Juan *et al.* (2010). This step is executed 10 times in order to have several feasible solutions.

4. Test each solution generated in the previous step with the original capacity of each vehicle *W* by using Monte Carlo Simulation. In this case, a route can be infeasible due to the capacity violation generated by the stochasticity of the demand, so the vehicle is allowed to return to the depot to re-load and continue the route. This will count as a route failure and will lead to an increase in the routing cost.

4. NUMERICAL EXPERIMENTS

To test the validity of our proposal, some tests were developed with three instances (20, 50, and 100 customers with 5 possible depot locations) used in Prins *et al.* (2006). The instances were modified in order to solve a single-depot location routing problem, the following assumptions were made: *(i)* there are not capacity constraints associated to depots; *(ii)* there are not fixed costs associated to the opening of depots; and *(iii)* the only cost to be considered is the one associated with distances. The algorithm was implemented as a Java application. The tests were executed in a 2.4GHz Core i5 MacBook with 8Gb RAM. An execution time of 10 seconds was established to determine the location of the depot. Table 1 summarizes the results of the location phase of the proposed method.

Instance	Depot to be opened	Solution Cost
Coord20-5-1	2	244.829
Coord50-5-1	1	937.696
Coord100-5-1	4	1654.743

Table 1. Summary of Depot Location.

Figure 1 shows each node of the instance Coord100-5-1, the blue dots correspond to customers, the red dots represent the closed depot locations and the yellow dot is the opened depot.

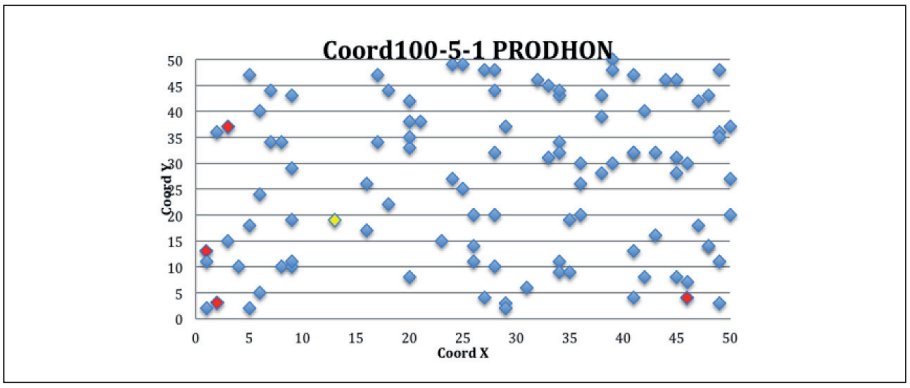


Figure 1. Graphical representation of 100 customer instance.

Nine different demand scenarios for each instance were generated to test the proposed approach. Each scenario was generated using a lognormal distribution with $\mu=E[D_i]$ *and* $Var[D_i]=0.2.d_i$ in order to have a medium variance level. Table 2 shows the results obtained, where *C* represents the routing cost of the solution, F represents the number of routes with demand greater than *W* (original capacity), i.e. the number of route failures, and *IC* represents the increased cost percentage due to stochastic demands. It can be clearly noticed that costs are only increased in a small percentage (12.97% in average in the worst case scenario) because of the uncertain demand.

Routing Phase	Coord20-5-1 Routes=3 W'=126			Coord50-5-1 Routes=13 W'=63			Coord100-5-1 Routes=27 W'=63		
	C	F	IC(%)	C	F	IC(%)	C	F	IC(%)
Base	229.526	0	0.00	910.255	0	0.00	1617.278	0	0.00
Scenario 1	229.526	0	0.00	1055.735	3	15.98	1677.556	1	3.72
Scenario 2	229.526	0	0.00	1047.717	2	15.10	1635.677	1	1.13
Scenario 3	248.099	1	8.09	910.255	0	0.00	1617.278	0	0.00
Scenario 4	276.170	1	20.32	910.255	0	0.00	1704.508	1	5.39
Scenario 5	276.170	1	20.32	910.255	0	0.00	1683.068	1	4.06
Scenario 6	276.170	1	20.32	910.255	0	0.00	1679.824	1	3.86
Scenario 7	276.170	1	20.32	910.255	0	0.00	1617.278	0	0.00
Scenario 8	245.849	1	7.11	959.604	2	5.42	1638.718	1	1.32
Scenario 9	276.170	1	20.32	1144.060	3	25.68	1731.010	2	7.03
Average			12.97			6.90			2.95

Table 2. Computational results for 9 different scenarios.

To test the effect of different variance levels, eight new scenarios were generated to test the solution for the 50-customer instance from Table 2, considering both small (*Var[Di]=0.05.di*) and small-to-median (*Var[Di]=0.10.di*) variances. Table 3 shows the validity of this approach when facing small and small-to-medium variance levels.

	Small variance level			Small-to-medium variance level		
	C	F	IC	C	F	IC
Scenario 1	910.255	0	0.00%	964.316	1	5.93%
Scenario 2	910.255	0	0.00%	910.255	0	0.00%
Scenario 3	910.255	0	0.00%	910.255	0	0.00%
Scenario 4	910.255	0	0.00%	910.255	0	0.00%
Scenario 5	910.255	0	0.00%	910.255	0	0.00%
Scenario 6	910.255	0	0.00%	910.255	0	0.00%
Scenario 7	910.255	0	0.00%	910.255	0	0.00%
Scenario 8	910.255	0	0.00%	910.255	0	0.00%

Table 3. Computational results for different variance levels.

In order to incorporate risk analysis to this work, 30 demand scenarios were generated to test the top-4 solutions obtained for the 50-customer instance with *Var[Di]=0.2.di*. It can be clearly noticed, in Figure 2, that SOL1 is the one with the lowest cost (908.6) but it is also the riskiest one, while SOL4 has a higher cost (911.3) but it is a low-risk solution. These two solutions are examples of decisions that could be taken by different decision-makers (DM), being the first one the typical decision of a risk-oriented DM and the latter the typical decision of a more conservative DM.

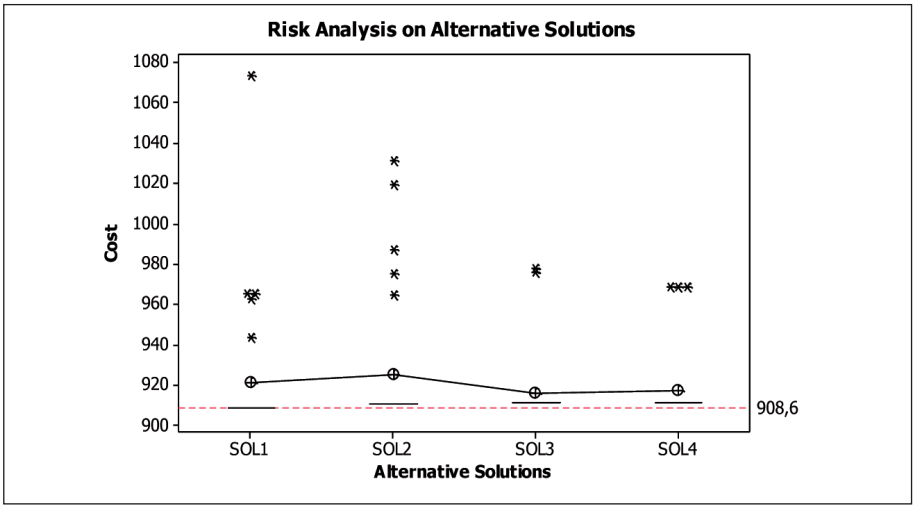


Figure 2. Risk Analysis.

5. CONCLUSIONS

A simulation-based algorithm was proposed to solve the Integrated Location Routing Problem considering uncertainty and risk factors. This method uses safety stocks and risk analysis to deal with the uncertainty of the demand. The validity of the method has been proved since only a small increase in cost is generated by the variability of the demand. It is especially strong when solving instances with small and small-to-medium variance levels. A similar approach could be used to tackle situations where other stochastic parameters are involved such as resources availability, stochastic customers, etc.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (grant TRA2013-48180-C3-3-P) and FEDER. Likewise we want to acknowledge the support received by the Department of Universities, Research & Information Society of the Catalan Government (Grant 2014-CTP-00001).

REFERENCES

Ahmadi-Javid, A., and Seddighi, A.H. (2013). "A Location-Routing Problem with Disruption Risk". *Transportation Research, Part E*, 53, 63-82.

Cordeau, J.F., Pasin,F., and Solomon, M.M. (2006). "An Integrated Model for Logistics Network Design". *Annals of Operations Research,* 144, (1), 59-82.

Herazo-Padilla, N., Montoya-Torres, J.R., Muñoz-Villamizar, A., Nieto-Isaza S., and Ramírez-Polo, L. (2013). "Coupling Ant Colony Optimization and DES to Solve a Stochastic Location-Routing Problem". *Proceedings of the 2013 Winter Simulation Conference*. Washington D.C. 3352-3362.

Juan, A., Faulin, J., Grasman, S., Riera, D., Marull, J., and Mendez, C. (2011). "Using Safety Stocks and Simulation to Solve the Vehicle Routing Problem with Stochastic Demands". *Transportation Research, Part C,* 19, 751-765.

Juan, A., Faulin, J., Ruiz, R., Barrios, B., and Caballé, S. (2010). "The SR-GCWS Hybrid Algorithm for Solving the Capacitated Vehicle Routing Problem". *Appl. Soft Comput.,* 10, (1), 215-224.

Muñoz-Villamizar, A., Montoya-Torres, J.R., Juan, A., and Cáceres-Cruz, J. (2013). "A Simulation-Based Algorithm for the Integrated Location and Routing Problem in Urban Logistics". *Proceedings of the 2013 Winter Simulation Conference. Washington D.C.,* 2032-2040.

Prins, C., Prodhon, C., and Wolfler-Calvo, R. (2006). "Solving the Capacitated Location-Routing Problem by a GRASP Complemented by a Learning Process and a Path Relinking". *4OR - A Quarterly Journal of Operations Research,* 4, (3), 221-238.

Prodhon, C., and Prins, C. (2014). "A Survey of Recent Research on Location-Routing Problems". *European Journal of Operational Research,* 238, 1-17.

Seyedhosseini, S.M., Bozorgi-Amiri, A., and Daraei, S. (2014). "An Integrated Location-Routing-Inventory Problem by Considering Supply Disruption". *iBusiness,* 6, 29-37.

78. CAPITAL ALLOCATION UNDER DEPENDENT RISKS: AN APPLICATION TO GHG EMISSIONS

L. Remuzgo, J.M. Sarabia and C. Trueba
University of Cantabria, Department of Economics, Spain,
{remuzgol, sarabiaj,truebac}@unican.es

ABSTRACT

In this study, we develop a framework to apply capital allocation principles to Greenhouse Gas (GHG) emissions. The approach relies on the use of the multivariate gamma distribution which allows establishing a formal model given its properties of dependence and positive skewness. We apply the presented resource allocation framework to the CO_2, CH_4 and N_2O emissions of the EU-28 in 2013. Using several allocation principles, we determine the incomes we should use for financing the reduction of the risks related to each GHG.

1. INTRODUCTION

The jointly effect of CO_2, CH_4 and N_2O emissions is very likely to have been the principal cause of global warming since the mid-20th century (IPCC, 2013). Despite the fact that the stabilization of the atmospheric concentration of GHGs is the ultimate goal of the United Nations Framework Convention on Climate Change (UNFCCC), such target does not attend the implied risks which are present in climate change (Obersteiner *et al.*, 2001). Thus, the UNFCCC also advocates the use of risk management measures (UNFCCC, 1992: Article 3.3).

Given the growing economic losses associated with climate risks, it is necessary to take gain from the background in risk management (IPCC, 2012). Thus, climate risk management attempts to apply financial risk management approaches to reduce

potential losses associated with climate change. In this context, we propose a resource allocation framework for determining the incomes we should use for financing the reduction of the risks related to each GHG, measured in terms of number of EU ETS allowances.

Despite the fact that allocation principles had been applied in other fields (Urbina and Guillén, 2014; Ai *et al.*, 2013), this is the first attempt to employ these techniques in the environmental area. In particular, we consider the CO_2, CH_4 and N_2O emissions of the EU-28 and calculate the number of EU ETS allowances we should intend for reduce the impact of each gas emissions. The methodological contribution of this study lies in the use of the multivariate gamma distribution as a formal model for the estimation of capital allocation principles, given its properties of dependence and positive skewness.

This paper is organized as follows. In Section 2 the multivariate gamma distribution is proposed as a model when there is dependence between risks. Considering the previous distribution, the expressions of different capital allocation principles are specified in Section 3. An application to GHG emissions is addressed in Section 4.

2. MODEL: MULTIVARIATE GAMMA DISTRIBUTION

The uncertainty around climate change consequences makes necessary the use of probability distribution functions which allow a better approach to considering a range of possible outcomes rather than a single most likely one (IPCC, 2004). In order to calculate different capital allocation principles, we choose the multivariate gamma distribution because it admits the existence of dependence between risks and it can be apply to positively skewed distributions. The class of gamma distributions is widely used in theoretical and applied actuarial statistics.

Let $Y_0, Y_1, \ldots, Y_m$ be mutually independent gamma random variables,i.e. $Y_i \sim Ga(\gamma_i, \alpha_i)$, where γ_i and α_i are the shape and scale parameters, respectively and let

$$X_i = \frac{\alpha_0}{\alpha_i} Y_0 + Y_i, \qquad i = 1, 2, \ldots, m. \tag{1}$$

Then, the joint distribution of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_m)^{\mathrm{T}}$ is a multivariate gamma distribution where the marginal distributions are standard gamma distributions but correlated (Mathai and Moschopoulus, 1991). Some important properties of this distribution are the following ones. The moment generating function of **X** is given by

$$M_{\underline{\mathbf{x}}}(t) = E\left[exp\left(Y_0 \sum_{i=1}^{m} \frac{\alpha_0}{\alpha_i} t_i\right)\right] \prod_{i=1}^{m} E[e^{t_i Y_i}] \\ = \left(1 - \sum_{i=1}^{m} \frac{t_i}{\alpha_i}\right)^{-\gamma_0} \prod_{i=1}^{m} \left(1 - \frac{t_i}{\alpha_i}\right)^{-\gamma_i}. \tag{2}$$

The next marginal information follows from the previous expression,

$$(i)\ X_i \sim Ga(\gamma_0 + \gamma_i, \alpha_i), \quad i = 1, 2, ..., m. \tag{3}$$

$$(ii)\ E(X_i) = \frac{\gamma_0 + \gamma_i}{\alpha_i}, \quad i = 1, 2, ..., m. \tag{4}$$

$$(iii)\ Var(X_i) = \frac{\gamma_0 + \gamma i}{\alpha_i^2}, \quad i = 1, 2, ..., m. \tag{5}$$

Finally, the covariance and the correlation coefficient are respectively:

$$Cov(X_i, X_j) = \frac{\gamma_0}{\alpha_i \alpha_j}, \quad i \neq j, \tag{6}$$

$$\rho(X_i, X_j) = \frac{\gamma_0}{\sqrt{(\gamma_0 + \gamma_i)(\gamma_0 + \gamma_j)}}, \quad i \neq j. \tag{7}$$

If $\gamma_0 \to 0$ then $X_i \sim \mathrm{Ga}(\gamma_i, \alpha_i)$, and the random vector **X** consists of m independent gamma random variables and, consequently, the joint distribution of **X** is a multivariate independent gamma distribution.

We focus on the case $m = 3$. The steps to estimate the parameters of a multivariate gamma distribution with three variables are the following.

Step 1: α_i parameters are estimated using (4) and (5):

$$\hat{\alpha}_i = \frac{E(X_i)}{Var(X_i)}, \qquad i = 1, 2, 3. \tag{8}$$

Step 2: γ_i parameters are estimated solving the equation system formed by the determinant of the covariance matrix (9) and (10):

$$\left|\sum\right| = \frac{\gamma_0(\gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3) + \gamma_1\gamma_2\gamma_3}{\alpha_1^2\alpha_2^2\alpha_3^2}, \tag{9}$$

$$\gamma_0 + \gamma_i = \frac{E(X_i)}{Var(X_i)}, \qquad i = 1, 2, 3. \tag{10}$$

where α_i, $i = 1, 2, 3$, had been estimated in the preceding step. The above procedure can be easily extended to m variables.

3. CAPITAL ALLOCATION PRINCIPLES USING THE MULTIVARIATE GAMMA DISTRIBUTION

Let $X_1, X_2, \ldots, X_m$ be m random losses arising from different risks which are materializing at a fixed future date T. Assume that each loss X_i has a finite mean and denote the cumulative distribution function (cdf) of Xi by $F_{X_i}(x)$. The aggregate loss is defined by,

$$S = \sum_{i=1}^{m} X_i, \qquad i = 1, \ldots, m. \tag{11}$$

Given an exogenous capital K to cover S, capital allocation seeks the distribution of K across the different risks, that is, to determine non-negative real numbers $(K_1, K_2, \ldots, K_m)$ satisfying the full allocation requirement:

$$\sum_{i=1}^{m} K_i = K. \tag{12}$$

Allocation principles are methods conceived to establish K_i considering different risk measures and assumptions. Thus, a constant allocation means dividing the total capital by the number of gases regardless of the contribution each gas emissions

makes to global climate change risk. Meanwhile, a proportional allocation gives more importance to the gases that represent a greater risk. Finally, and optimal allocation takes into account the contribution of each gas considering the dependence between the different emissions within the climate system (Dhaene *et al.*, 2012).

Name	K_i, $i = 1, \ldots, m$
Haircut	$\frac{K}{\sum_{i=1}^{m} F_{X_i}^{-1}(p)} F_{X_i}^{-1}(p)$
SD Esscher	$\frac{K}{E[S] + a\sum_{i}^{m} \sigma_{X_i}} [E(X_i) + a\sigma_{X_i}], \quad a \geq 0$ $\frac{K}{\sum_{i=1}^{m} E\left[\frac{X_i e^{aX_i}}{E(e^{aX_i})}\right]} E\left[\frac{X_i e^{aX_i}}{E(e^{aX_i})}\right], \quad a > 0$
Covariance Overbeck I	$\frac{K}{Var[S]} Cov(X_i, S)$ $\frac{K}{E[S] + \sum_{i=1}^{m} \sigma_{X_i}} [E(X_i) + \frac{a}{\sum_{i=1}^{m} \sigma_{X_i}} Cov(X_i, S)], \quad a \geq 0$

Table 1. Capital allocation principles.

We refer to Table 1 for a summary of different capital allocation principles, considering that the losses follow a multivariate gamma distribution.

4. AN APPLICATION TO THE MAIN GHG EMMISSIONS

In this section, we apply the allocation principles estimated in the previous section to per-capita CO_2, CH_4 and N_2O emissions in order to calculate the incomes we should use for financing the reduction of the risks related to each one. Such emissions, measured in gigagrams of CO_2-eq, have been taken from the estimates compiled by the European Environment Agency in the approximated EU GHG inventory for 2013. Given that the EU has planned to use a share of the revenues from the auctioning of EU ETS allowances for climate and energy related purposes (European Commission, 2014), the incomes are given in terms of the number of these allowances in 2013, that is, 2039153 gigagrams of CO_2-eq. We refer to Figure 1 for a

graphical representation of the positive skew of the per-capita GHG emissions. Table 2 shows the distribution of EU ETS allowances among the three gases based on different capital allocation principles.

Based on the information of the GHG emissions from 2008 to 2013 and using a simulation process, we study the sensitivity of K_1 considering the future time evolution of its variance under the different criteria. The evolution of the variances is consistent with the reduction of the emissions, which is in turn associated with the decrease of allowances, for all the criteria but the Covariance principle (see Figure 2).

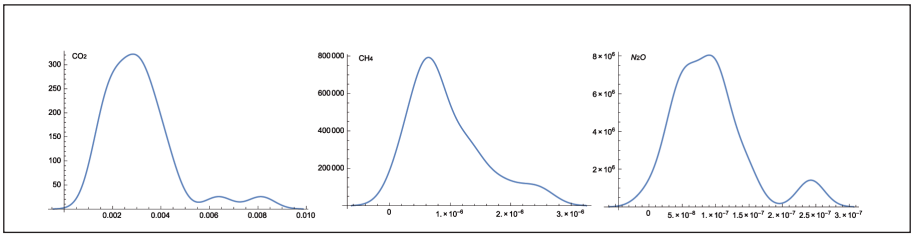


Figure 1. Density function of CO_2, CH_4 and N_2O emissions (2013).

Name	K_1	K_2	K_3
Haircut	2038378	708	67
SD	2038427	663	63
Esscher	2038474	619	60
Covariance	2038955	146	52
Overbeck I	2038565	530	58

Table 2. Capital allocation principles.

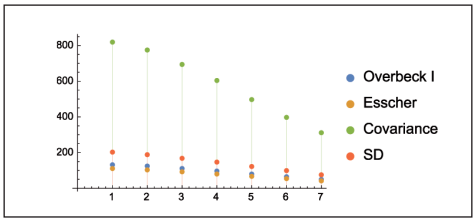


Figure 2. Evolution of the variance of CO_2 EU ETS allowances.

ACKNOWLEDGMENTS

The authors thank to Ministerio de Economía y Competitividad (Project ECO 2013-48326-C2-2-P) and Ministerio de Educación, Cultura y Deporte (FPU 13/02155) for the partial support of this work.

REFERENCES

Ai, J., Brockett, P.L., Golden, L.L., and Guillén, M. (2013). "A Robust Unsupervised Method for Fraud Rate Estimation". *Journal of Risk and Insurance*, 80, 121-143.

Bühlmann, H. (1970). *Mathematical Methods in risk Theory*. Berling: Springer-Verlag.

Dhaene, J., Tsanakas, A., Valdez, E.A., and Vanduffel, S. (2012). "Optimal Capital Allocation Principles". *Journal of Risk and Insurance*, 79, 1-28.

European Commission (2014). *Report from the Commission to the European Parliament and the Council on Progress towards Achieving the Kyoto and EU 2020 Objectives*.

Gerber, H.U. (1981). "The Esscher Premium Principle: A criticism Comment". *ASTIN Bulletin*, 12, 139-140.

Intergovernmental Negotiating Committee for a Framework Convention on Climate Change (1992). *UNFCCC of the Second Part of Its Fifth Session*, celebrated in New York from 30 April to 9 May 1992. United Nations, New York.

IPCC (2004). *IPCC Workshop on Describing Scientific Uncertainties in Climate Change to Support Analysis of Risk and of Options.* A Special Report of WG I of the IPCC [Manning, M., Petit, M., Easterling, D., Murphy, J., Patwardhan, A., Rogner, H.-H., Swart, R., Yohe, G., (eds.)]. IPCC, Geneva, Switzerland.

IPCC (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. A Special Report of WG I and II of the IPCC [Field, C.B.,

Barros, V., Stocker, T.F., Qin, D., Dokken, D.J., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.-K., Allen, S.K., Tignor, M., Midgley, P.M. (eds.). Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.

IPCC (2013). *Climate Change 2013: The Physical Science Basis*. Contribution of WP I to the 5 AR of the IPCC [Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex V., Midgley P. M. (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, 1535 pp.

Mathai, A.M., and Moschopoulos, P.G. (1991). "On a Multivariate Gamma". *Journal of Multivariate Analysis*, 39, 135-153.

Obersteiner, M., Azar, C., Kauppi, P.E., Moellersten, K., Moreira, J., Nilsson, S., Read, P., Riahi, K., Schlamadinger, B., Yamagata, Y., Yan, J., and Van Ypersele, J.P. (2001). *Managing Climate Risk*. IIASA Interim Report, IR-01-051.

Urbina, J., and Guillén, M. (2014). "An Application of Capital Allocation Principles to Operational Risks and the Cost of Fraud". *Expert Systems with Applications*, 41, 7023-7031.

79. A TEST ON SEPARATE FAMILIES OF HYPOTHESES: DISTINGUISH BETWEEN UNIFORM AND EXPONENTIAL DISTRIBUTIONS

P. Rochet
Université de Nantes, Laboratoire de Mathématiques Jean Leray, France
paul.rochet@univ-nantes.fr

I. Serra
Centre de Recerca Matemàtica, Barcelona, Spain
iserra@crm.cat

ABSTRACT

In extreme value theory, the performances of most estimators of the tail index highly rely on the family of the limit distribution (e.g. Weibull, Gumbel or Fréchet). In this paper, we consider the test on separate families of hypothesis for exponential versus uniform distributions in order to determine the family of the limit distribution. We derive the most powerful test for such hypothesis under mild restrictions. In particular, we show that the optimal test in this framework is not the likelihood ratio test.

INTRODUCTION

The starting point of Extreme Value Theory (EVT) was characterized by two fundamental theorems, Fisher-Tippet-Gnedenko's theorem and Pickands-Balkema-De Haan's theorem. From practical point of view, they characterize the tail distribution defined by the extremes values with a real number, called the tail index and denoted by ξ, see McNeil *et al*. (2005), Beirlant *et al*. (2004), Finkenstadt and Rootzén (2003), Coles (2001) and Embrechts *et al*. (1997). While these theorems provide an approach

to estimate the tail index, the produced results are not satisfactory in many situations. For instance, several authors have highlighted the estimation problems that arise from the second approach, see Hosking and Wallis (1987), Castillo and Hadi (1997), Zhang and Stephens (2009).

The second fundamental theorem have widened the use of the Generalized Pareto Distribution (GPD) in extreme values theory as a model for tails. According to the behavior of the probability density functions of the GPD, we can distinguish three submodels that are separated by the exponential distribution ($\xi = 0$) and the uniform distribution ($\xi = -1$). Summarizing, for $\xi > 0$ the Pareto submodel, for $\xi = 0$ the exponential distribution, for $-1 < \xi < 0$ a family of decreasing probability densities functions with finite support, for $\xi = -1$ the uniform distribution and for $\xi < -1$ a family of increasing densities with finite support. The study highlights the problem of misspecification of the submodel, see Castillo and Serra (2015).

Therefore, from a statistical modelling point of view, the GPD has to be considered as three models separated by the exponential and uniform distribution, respectively. Since no method is able to ensure the correct classification between these three (see Castillo and Serra, 2015), the estimation of parameters should be made subsequently to the choice of model and even more for small samples. Of course, the first recommendation is to specify, a priori, the submodel of the GPD in order to estimate the parameters. This depends on the dataset and would have to be decided from the framework, which is not always possible. Therefore, a test to distinguish between exponential and uniform distribution is proposed in order to contribute to the correct classification of the tail.

Tests on separate families of hypotheses have been originally studied in Cox (1961) who proposed a generalization of the Neyman-Pearson principle, based on a likelihood ratio criterion. While the likelihood ration test is not necessarily optimal, it provides a methodology to test a wide range of hypotheses. Examples studied in the literature include an invariant test of exponential versus normal or uniform distributions in Uthoff (1970), normal versus Cauchy distribution in Franck (1981) and so on.

The test on separate families of hypothesis for exponential versus uniform distributions proposed is the most powerful test for such hypothesis under mild restrictions. The derived statistic to realize the contrast is

$$\frac{\sum_{i=1}^{n} X_i}{X_{(n)}} - 1$$

where the sample is denoted by $\{X_i\}$ and $X_{(n)}$ correspond to the maximum. That is directly related with the properties of the tail of this distributions, see Villaseñor-Alva and González-Estrada (2009), in front of other tests between this families, see Uthoff (1970). Finally, it is interesting to remark that this test is the optimal test in this framework and it is not the likelihood ratio test.

2. UNIFORM *vs* EXPONENTIAL DISTRIBUTION TEST

Let $X_1, \ldots, X_n$ be an iid sample from a distribution μ. Assume that we know that μ is either uniform on an interval $[0, \theta]$ for some $\theta > 0$ (noted $\mathcal{U}(\theta)$), or exponential with parameter $\lambda > 0$ (noted $E(\lambda)$). We consider the test on the hypothesis $\mathcal{H}_0 : \exists \theta > 0, \mu \sim \mathcal{U}(\theta)$, the alternative being $\mathcal{H}_1 : \exists \lambda > 0, \mu \sim \mathcal{E}(\lambda)$. The test hypotheses can be written as

$$\mathcal{H}_0 = \cap_{\theta>0} \mathcal{H}_{0,\theta} \text{ with } \mathcal{H}_{0,\theta} : \mu \sim \mathcal{U}(\theta)$$

and

$$\mathcal{H}_1 = \cap_{\lambda>0} \mathcal{H}_{1,\lambda} \text{ with } \mathcal{H}_{1,\lambda} : \mu \sim \mathcal{E}(\lambda).$$

Definition 2.1. A test $\Phi(X)$ has constant level α if for all distributions μ under $\mathcal{H}_0$, $\mathbb{E}_\mu(\Phi(X)) = \alpha$. Similarly, a test has constant power $1 - \beta$ if for all distribution μ under $\mathcal{H}_1$, $\mathbb{E}_\mu(\Phi(X)) = 1 - \beta$.

Definition 2.2. We say that a test $\Phi(X)$ is homogenous if $\Phi(tX) = \Phi(X)$ for all $t > 0$.

A test is homogenous if it can be expressed as a function of the normalized sample $X/\| X \|$ for any norm $\| . \|$ on $\mathbb{R}^n$. Equivalently, writing the test as $\Phi(X) = \mathbf{1}\{X \in \mathcal{A}\}$, for $\mathcal{A}$ a Borel set of $\mathbb{R}^n$, a test is homogenous if, and only if, A is a cone of $\mathbb{R}^n$.

The reason for using an homogenous test in this framework is clear: since both the uniform and exponential distribution are stable under multiplication by a positive number, the level and power of an homogenous test are constant. Indeed, if X is uniformly distributed on an interval $[0, \theta]$ (resp. X has exponential distribution) then tX (with $t > 0$) remains uniformly distributed (resp. exponentially distributed) so that the distribution of $X / \| X \|$ does not depend on the parameter.

Let $L_0(\theta, X) = \theta^{-n} 1\{X_i \leq \theta, i = 1, \ldots, n\}$ denote the likelihood function under $\mathcal{H}_0$ and $L_1(\lambda, X) = \lambda^n \exp(-\lambda \sum_{i=1}^n X_i)$ the likelihood under $\mathcal{H}_1$. Based on the Neyman-Pearson likelihood ratio test theory, a somewhat natural approach would be to reject $\mathcal{H}_0$ as soon as the maximum likelihood ratio exceeds a certain threshold k_α. In this particular case, the maximum likelihood ratio is given by

$$\frac{\sup_{\lambda>0} L_1(\lambda, X)}{\sup_{\theta>0} L_0(\theta, X)} = \left(\frac{n}{e} \frac{\sum_{i=1}^n X_i}{X_{(n)}} \right)^n,$$

with $X_{(n)} = \max\{X_1, \ldots, X_n\}$, so that the test can be reduced to rejecting $\mathcal{H}_0$ as soon as

$$T(X) := \frac{\sum_{i=1}^n X_i}{X_{(n)}} - 1$$

exceeds a suitable threshold. Observe that this test is homogenous and therefore, has constant level and power. However, while this test is based on the Neyman-Pearson criterion, nothing indicates that it is the most powerful test. Indeed, the Neyman-Pearson Lemma tells us that such test is the most powerful only if $\mathcal{H}_0$ and $\mathcal{H}_1$ contain only one element. Surprisingly, we show in the next theorem that the most powerful test in this case is obtained by taking the opposite rejection region.

Theorem 2.1. The test $\Phi = 1\{T(X) \leq c_\alpha\}$ with $c\alpha$ such that $\mathbb{P}(T(X) \leq c_\alpha | \mathcal{H}_0) = \alpha$ is the most powerful homogenous test of level $\alpha \in (0, 1)$ to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

ACKNOWLEDGMENTS

Research expenses were founded by projects MTM2012-31118 from Spanish MINECO and La Caixa foundation.

REFERENCES

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics.

Castillo, E., and Hadi, A.S. (1997). "Fitting the Generalized Pareto Distribution to Data". *J. Amer. Statist. Assoc*., 92, 1609-1620.

Castillo, J., and Serra, I. (2015). "Likelihood Inference for Generalized Pareto Distribution". *Computational Statistics and Data Analysis*, 83, 116-128.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

Cox, D.R. (1961). "Tests of Separate Families of Hypotheses". *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105-123 (June).

Cox, D.R. (1962). "Further Results on Tests of Separate Families of Hypotheses". *Journal of the Royal Statistical Society*. Series B (Methodological), 406-424.

Embrechts, P., Klüppelberg, C., and Mikosch., T. (1997). "Modelling Extremal Events for Insurance and Finance". Springer, *Applications of Mathematics*, 33.

Finkenstädt, B., and Rootzén, H. (2003). "Extreme Values in Finance, Telecommunications and the Enviroment". *Monographs on Statistics and Applied Probability*, 99.

Franck, W.E. (1981). "The Most Powerful Invariant Test of Normal versus Cauchy with Applications to Stable Alternatives". *Journal of the American Statistical Association*, 76, 1002-1005.

Hosking, J.R.M., and Wallis, J.R. (1987). "Parameter and Quantile Estimation for the Generalized Pareto Distribution". *Technometrics*, 29, 339-349.

McNeil, A.J., Frey, R., and Embrechts P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

Pace, L., and Salvan, A. (1990). "Best Conditional Tests for Separate Families of Hypotheses". *Journal of the Royal Statistical Society*. Series B (Methodological), 125-134.

Pereira, B. D. B. (1977). "A Note on the Consistency and on the Finite Sample Comparisons of some Tests of Separate Families of Hypotheses". *Biometrika*, 64, 109-113.

Uthoff, V.A. (1970). "An Optimum Test Property of Two Well-Known Statistics". *Journal of the American Statistical Association*, 65, 1597-1600.

Villaseñor-Alva, J. A., and González-Estrada, E. (2009). "A Bootstrap Goodness of Fit Test for the Generalized Pareto Distribution". *Computational Statistics and Data Analysis*, 53, 3835-3841.

Zhang, J., and Stephens, M. A. (2009). "A New and Efficient Estimation Method for the Generalized Pareto Distribution". *Technometrics*, 51, 316-325.

80. MODELING *N* ELECTRICAL COMPONENT FAILURES USING NON-STATIONARY MARKOVIAN ARRIVAL PROCESSES

J. Rodríguez
Universidad Carlos III de Madrid, Department of Statistics, Spain
joannavirginia.rodriguez@uc3m.es

R. E. Lillo
Universidad Carlos III de Madrid, Department of Statistics, Spain
rosaelvira.lillo@uc3m.es

P. Ramírez-Cobo
Universidad de Cádiz, Department of Statistics
and Operational Research, Spain
pepa.ramirez@uca.es

ABSTRACT

The matricial formulation of the Markovian arrival process (*MAP*) allows for an analytic tractability of the process enabling the calculation of significant quantities and probabilities of interest associated to the system, therefore the risk of failure can be measured both in the stationary case and at any time t.

In this article, the non-stationary *MAP* is considered to model the failures of N electrical components that are considered to be structurally equal, for which it is not reasonable to assume that the inter-failure times related to each component are independent nor identically distributed. We implement a moments matching based estimation approach to fit a real data set provided by the Spanish electrical group Iberdrola to a non-stationary *MAP*.

1. INTRODUCTION

Reliability in this context is described as the probability of a system or a component to function under stated conditions for a specified period of time. Component failures are caused by faults or errors in the components, that can be caused by internal (due to the internal structure of the system) or external (environmental condition) factors. As a component may fail a repair or a replacement may take place so that the component goes back to functioning as soon as possible. In the literature, when the lifetime and deterioration of a system is submitted to external failures, they are called shocks. Shock and wear models are studied in reliability for modeling systems under environmental conditions. Reliability and maintenance policies of systems considering random component failure have been widely studied in the literature. However, the most common assumption is that the failures arrive independently, considering that they arrive according to Poisson, renewal, or PH renewal processes. Considering that the inter-failure times are not independent, the most appropriate type of process to model the arrivals are the Markovian arrival processes (*MAP*). The *MAP* is an important class of point processes that generalizes the Poisson arrival process, it allows for dependent and non-exponential interarrival times. The matricial formulation of the *MAP* allows for an analytic tractability of the process enabling the calculation of significant quantities and probabilities of interest associated to the system, therefore the risk of failure can be measured both in the stationary case and at any time t. The *MAP* has been used in the literature to model the arrival of failures in shock and wear models, for example, see Montoro-Cazorla *et al*. (2009, 2012, 2014) and Pérez-Ocón and Segovia (2009).

As stated, the *MAP* is a suitable point process that properly fits non-Poisson processes, allowing for dependent and non-exponentially distributed interfailures times. In its stationary version, the *MAP* assumes equally distributed inter-failures times, but in the non-stationary counterpart, the process which shall be considered in this chapter, this assumption is relaxed. The matricial formulation of the *MAP* allows for an analytic tractability of the process enabling the calculation of significant quantities and probabilities of interest associated to the reliability of the system, such as the cumulative distribution of the inter-failure times, the probabilities and expected number of failures at a specific time t.

We present a system composed of N components, where the failures for each component are repaired in a time that is far shorter than the operational times, and are considered as instant repairs. When the system does not work properly, a replacement can also take place. As stated before, we cannot assume independence between failures, also as a replacement or repair takes place, there is no evidence that the underlying distribution is maintained, meaning that we are not dealing with renewal processes, considering these properties, it will be assumed that the failures follow a two-state Markovian arrival process (MAP_2), and to allow more versatility, we consider the nonstationary version of the process.

We present some theoretical properties such as identifiability and an estimation procedure, to obtain the parameters of the nonstationary *MAP* model. Identifiability means the uniqueness of the parametrization of the model, if this property does not hold, the standard estimation methods will fail to provide a suitable representation of the process. The Markovian arrival process allows for a better tractability of the process allowing the calculation of important quantities and probabilities of interest associated to the reliability of the system, such as the probability and mean number of failures in an interval $(0, t]$.

The article is organized as follows. In Section 2 we recall some definitions and important results regarding the *MAP* and we present our estimation results in Section 3.

2. THE *MAP*$_2$ AND SOME IMPORTANT RESULTS

In this section, we give briefly a review of the two-state *MAP* and prove that the nonstationary *MAP* is a nonidentifiable process, meaning that it will lack a unique representation, due to this fact, we will follow the results of Bodrog *et al.* (2008), and provide a canonical representation for the nonstationary MAP_2 in terms of a reduced number of parameters.

The MAP_2 is a doubly stochastic process $\{J(t), \mathrm{N}(t)\}$, where $J(t)$ represents an irreducible, continuous, Markov process with state space $S = \{1, 2\}$.

The counting process $\{N(t), t \geq 0\}$ represents the number of failures (or repairs) in the interval $(0, t]$ of the (identically distributed) constituting components. For a thorough definition of the MAP_2, see Ramírez-Cobo *et al.* (2010, 2012). Concerning the general m-state *MAP*, we refer the reader to Neuts (1979) and Lucantoni *et al.* (1990).

A non-stationary MAP_2 will be represented by $\mathcal{M} = \{\boldsymbol{\alpha}, D_0, D_1\}$ where

$$\boldsymbol{\alpha} = (\alpha,\ 1-\alpha), \quad D_0 = \begin{pmatrix} x & y \\ z & u \end{pmatrix}, \quad D_1 = \begin{pmatrix} w & -x-y-w \\ v & -z-u-v \end{pmatrix}, \tag{1}$$

and without loss of generality it is assumed that $u \leq x$. Where

$$x = -\lambda_1, \quad y = \lambda_1 p_{120}, \quad w = \lambda_1 p_{111},$$
$$z = \lambda_2 p_{210}, \quad u = -\lambda_2, \quad v = \lambda_2 p_{211}.$$

Let T_k be the random variable representing the time between the $(k-1)$-th and the k-th failures. T_k is phased-type distributed with representation $\{\alpha_k, D_0\}$, where

$$\boldsymbol{\alpha}_k = \boldsymbol{\alpha} \left(P^{\star}\right)^{k-1},$$

which implies that the inter-failure times are not identically distributed. Then the cumulative distribution function and moments of T_k are given by

$$F_{T_k}(t) = 1 - \boldsymbol{\alpha}_k e^{D_0 t}\mathbf{e}, \tag{2}$$
$$\mu_{k,m} = E\left(T_k^m\right) = m!\boldsymbol{\alpha}_k \left(-D_0\right)^{-m}\mathbf{e},$$

for $m = 1, 2, \ldots$ and where e represents a unit vector.

The identifiability is an important property when estimation of the process parameters is considered. If there exist multiple representations of the same process, this will imply infinite solutions and possibly non-convergence of the typical maximum likelihood approaches. In practice, it is usual that only the inter-failure times are observed when inference of the process is considered. Therefore, our definition of identifiability is stated in terms of the inter-failure times distribution.

Definition 2.1. The non-stationary MAP_2 is a nonidentifiable process if for any fixed MAP_2 with representation $\mathcal{M}$, then there exists another MAP_2 with different representation $\widetilde{\mathcal{M}}$ such that

$$(T_1, \ldots, T_n) \stackrel{d}{=} \left(\widetilde{T}_1, \ldots, \widetilde{T}_n\right), \quad \textit{for all } n \geq 1, \tag{3}$$

where Ti represent the time between the (i–1)-th and i-th event occurrences in the MAP_2 defined by $\mathcal{M}$ (similarly define $\widetilde{T}_i$).

In the next results we prove the nonidentifability of the non-stationary *MAP*; that is, equalities (3) holds for all $n \geq 1$. And the canonical representation for the non-stationary MAP_2.

Theorem 2.1. The non-stationary MAP_2 is a nonidentifiable process.

Theorem 2.2. The canonical representation for the non-stationary MAP_2 is unique and given by

$$\boldsymbol{\alpha}^c = (\alpha_0,\ 1-\alpha_0), \quad D_0^c = \begin{pmatrix} x_0 & y_0 \\ 0 & u_0 \end{pmatrix}, \quad D_1^c = \begin{pmatrix} -x_0 - y_0 & 0 \\ v_0 & -u_0 - v_0 \end{pmatrix}, \tag{4}$$

if $\gamma > 0$. On the other hand, for those non-stationary MAP_2s such that $\gamma \leq 0$, then the canonical form is

$$\boldsymbol{\alpha}^c = (\alpha_0,\ 1-\alpha_0), \quad D_0^c = \begin{pmatrix} x_0 & y_0 \\ 0 & u_0 \end{pmatrix}, \quad D_1^c = \begin{pmatrix} 0 & -x_0 - y_0 \\ -u_0 - v_0 & v_0 \end{pmatrix}, \tag{5}$$

where $u_0 \leq x_0 < 0, y_0, v_0 \geq 0, x_0 + y_0 \leq 0$ and $u_0 + v_0 \leq 0$. The initial probability α_0 is defined as

$$\alpha_0 = \frac{-x_0 - x_0 u_0 r_1}{u_0 - y_0 - x_0}, \tag{6}$$

where $r_1 = \alpha(-D_0)^{-1}$ e.

3. PARAMETER ESTIMATION ALGORITHM

In this section we introduce the modeling framework based on a set of operational times $\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)}$ as in (7). Consider an N electrical component framework, each of them subject to failures. After each failure, the component is instantly repaired, and we record the consecutive operational times of each component. The aim is to estimate the performance of the system on base on N real sequences of such operational times, $\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)}$, where

$$\begin{aligned} \mathbf{t}^{(1)} &= \left(t_1^{(1)}, t_2^{(1)}, \ldots, t_{n_1}^{(1)}\right), \\ \mathbf{t}^{(2)} &= \left(t_1^{(2)}, t_2^{(2)}, \ldots, t_{n_2}^{(2)}\right), \\ &\vdots \\ \mathbf{t}^{(N)} &= \left(t_1^{(N)}, t_2^{(N)}, \ldots, t_{n_N}^{(N)}\right), \end{aligned} \tag{7}$$

where ni denotes the size of the sample $\mathbf{t}^{(i)}$, for $i = 1, \ldots, N$. We assume that the N components are identical and the sequences of operational times $\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)}$, are independent among them.

In this case, let T_k be the random variable representing the operational time between the $(k-1)$-th and k-th failure of any given component. Unlike classical model assumptions, we cannot assume that the random variables $\{T_k\}_{k\geq 1}$ are uncorrelated, and then, they cannot be considered independent. Also, the random variables $\{T_k\}_{k\geq 1}$ are not necessarily identically distributed.

The parameters of the canonical representation for the non-stationary MAP_2 are estimated following Rodríguez *et al.* (2015), where a moments matching based approach is proposed.

The electrical failure times provided by the Spanish private electrical utility company, Iberdrola, are collected from an electrical framework composed of 926 components that supply energy. The times in which the components fail are recorded. If an electrical component or generator fails it is repaired and restarted. The repair times are considered negligible in comparison with the operational times.

The data presents an structure as in (7), where

$$\begin{aligned} &N = 926, \\ &\min\{n_1, \ldots, n_N\} = 1, \\ &\max\{n_1, \ldots, n_N\} = 42, \end{aligned} \tag{8}$$

where n_i, for $i = 1, \ldots, N$, denotes the number of observed operational times of the i-th component.

The sample moments needed for estimating the model are

$$(\overline{\mu_{1,1}},\ \overline{\mu_{1,2}},\ \overline{\mu_{1,3}},\ \overline{\mu_{2,1}},\ \overline{\mu_{3,1}}) = (79.226, 7.478{\times}10^3, 7.2911{\times}10^3, 69.0582, 67.5977)$$

The inference approach described provided the following estimates of the non-stationary MAP_2 in the first canonical form:

$$\hat{\boldsymbol{\alpha}}^1 = (0.4608, 0.5392), \quad \hat{D}_0^1 = \begin{pmatrix} -0.2394 & 0.1345 \\ 0 & -0.0104 \end{pmatrix}, \quad \hat{D}_1^1 = \begin{pmatrix} 0.1049 & 0 \\ 0.0067 & 0.003 \end{pmatrix}$$

with estimated moments given by

$$(\widehat{\mu_{1,1}},\ \widehat{\mu_{1,2}},\ \widehat{\mu_{1,3}},\ \widehat{\mu_{2,1}},\ \widehat{\mu_{3,1}}) = (78.8950, 7.535{\times}10^3, 7.2633{\times}10^3, 69.0864, 67.5712$$

and objective function equal to $\delta_\tau^1 = 9.0324 \times 10^{-5}$.

On the other hand, the estimated MAP_2 parameters under the second canonical form were

$$\hat{\boldsymbol{\alpha}}^2 = (0.8207, 0.1793), \quad \hat{D}_0^2 = \begin{pmatrix} -0.0104 & 0.0104 \\ 0 & -16.5378 \end{pmatrix}, \quad \hat{D}_1^2 = \begin{pmatrix} 0 & \\ 11.7651 & 4.7 \end{pmatrix}$$

with estimated moments given by

$$(\widehat{\mu_{1,1}},\ \widehat{\mu_{1,2}},\ \widehat{\mu_{1,3}},\ \widehat{\mu_{2,1}},\ \widehat{\mu_{3,1}}) = (78.7930, 7.5583{\times}10^3, 7.2513{\times}10^3, 68.3119, 68.311$$

and objective function $\delta_\tau^2 = 4.0324 \times 10^{-4}$. The log-likelihoods were

$$\log f(\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)} | \hat{D}_0^1, \hat{D}_1^1) = -5.3790{\times}10^4, \quad \log f(\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)} | \hat{D}_0^2, \hat{D}_1^2) = -5.$$

which provides evidence in favor of the estimate $\{\hat{\alpha}^1, \hat{D}_0^1, \hat{D}_1^1\}$.

ACKNOWLEDGMENTS

The research of the authors is supported by Grants ECO2011-25706, ECO2012-38442 and MTM2012-36163 (Spanish Ministry of Science and Innovation), P11-FQM-7603 and FQM-329 (Junta de Andalucía, Spain). The third author was supported by Consolider "Ingenio Mathematica" through her post-doc contract.

REFERENCES

Bodrog, L., Heindlb, A., Horvátha, G., and Telek M. (2008). "A Markovian Canonical Form of Second-Order Matrix-Exponential Processes". *European Journal of Operational Research*, 190, 459-477.

Lucantoni D, Meier-Hellstern K, Neuts MF. (1990). "A Single Server Queue with Server Vacations and a Class of Nonrenewal Arrival Processes". *Advances in Applied Probability*, 22, 676-705.

Montoro-Cazorla, D., and Pérez-Ocón R. (2012). "A Shock and Wear System under Environmental Conditions Subject to Internal Failures, Repair, and Replacement". *Reliability Engineering and System Safety*, 99, 55-61.

Montoro-Cazorla, D., and Pérez-Ocón R. (2014). "Matrix Stochastic Analysis of the Maintainability of a Machine under Shocks". *Reliability Engineering and System Safety*, 121, 11-17.

Montoro-Cazorla, D., Pérez-Ocón R., and Segovia MC. (2009). "Replacement Policy in a System under Shocks Following a Markovian Arrival Process". *Reliability Engineering and System Safety*, 94, 497-502.

Neuts MF. (1979). "A Versatile Markovian Point Process". *Journal of Applied Probability*, 16, 764-779.

Pérez-Ocón R., and Segovia MC. (2009) "Shock Models under a Markovian Arrival Process". *Mathematical and Computer Modeling*, 50, 879-884.

Ramírez-Cobo, P., and Lillo, R.E. (2012). "New Results About Weakly Equivalent MAP_2 and MAP_3 Processes". *Methodology and Computing in Applied Probability*, 14, 421-444.

Ramírez-Cobo, P., Lillo, R.E., and Wiper, M. (2010). "Non Identifiability of the Two-State Markovian Arrival Process". *Journal of Applied Probabability*, 46, 630-649.

Rodríguez, J., Lillo, R.E. and Ramírez-Cobo, P. (2015). "Failure Modeling of an Electrical N-Component Framework by the Non-Stationary Markovian Arrival Process". *Reliability Engineering and System Safety*, 134, 126-133.

81. LA DISTRIBUCIÓN PARETO POSITIVE STABLE, UN MODELO ESTOCÁSTICO PARA EVENTOS EXTREMOS

A. J. Sáez-Castillo
Universidad de Jaén, Departamento de Estadística e Investigación Operativa, España
ajsaez@ujaen.es

J. M. Sarabia y Faustino Prieto
Universidad de Cantabria, Departamento de Economía, España

RESUMEN

Las distribuciones que se utilizan como modelos de eventos extremos desempeñan un papel fundamental en la cuantificación del riesgo asociado a dichos eventos. Presentamos un estudio heurístico que muestra la posibilidad de incluir la distribución Pareto Positive Stable como uno de estos modelos. Los resultados ponen de manifiesto que puede proporcionar mejores ajustes que otras distribuciones comúnmente utilizadas en la modelización de fenómenos naturales, y que también puede proporcionar buenos ajustes sobre las pérdidas ocasionadas por éstos. El manejo de la distribución mediante el *software* estadístico R a través de la librería ParetoPosStable, que incluye tanto las funciones más relevantes (de densidad, de distribución, cuantil) como la generación de valores aleatorios, y que resuelve el problema de la estimación y la valoración de la bondad del ajuste, facilitan su aplicación en problemas reales. Como conclusión, creemos que esta distribución puede ser utilizada en el cálculo de medidas del riesgo como el valor de retorno para períodos de retorno dados, el Value at Risk o el Tail Value at Risk.

1. INTRODUCCIÓN

Frente a la mayoría de los métodos estadísticos clásicos, orientados mayoritariamente a analizar lo que ocurre en el centro de una distribución estadística, la teoría de eventos extremos se ocupa de las colas de dicha distribución o, dicho de otra forma, de sus valores más extremos. El riesgo asociado a estos valores extremos ha de tenerse en cuenta en todas las áreas del manejo del riesgo. De hecho, uno de los grandes desafíos en este campo es la implementación de una metodología de modelado estadístico que permita la medida de sus probabilidades de ocurrencia y, sobre todo, de sus consecuencias.

El presente trabajo pretende comparar el ajuste, mediante la distribución Pareto Positive Stable (PPS), de datos procedentes de diferentes contextos, donde alguna variable cuantifica la ocurrencia de un evento extremo, con el que proporcionan otras distribuciones habitualmente utilizadas en la literatura. Valoraremos la precisión de cada distribución en los ajustes y cuantificaremos medidas del riesgo asociadas a ellos.

2. EVENTOS EXTREMOS Y MEDIDAS DEL RIESGO

Dada una muestra $X_1, X_2, ..., X_n$ de variables aleatorias independientes e idénticamente distribuidas con función de distribución $F(x) = Pr(X \leq x)$, que representan valores pasados de una variable X asociada a la ocurrencia de un fenómeno natural (precipitaciones, caudales, vientos, temperaturas...) o a los daños ocacionados por tales fenómenos en un determinado período de tiempo, se considera un riesgo o evento extremo el hecho de que la variable supere un determinado valor, x.

Una forma de definir el evento extremo es asociarlo a su frecuencia de ocurrencia. Así, si consideramos un evento extremo como aquel que ocurre, en promedio, una vez cada T instantes de tiempo, entonces, la probabilidad de que ocurra en el siguiente período de tiempo es $Pr(X > x_T) = 1 - F(x_T) = 1/T$, y, por tanto, $x_T = F^{-1}\left(1 - \frac{1}{T}\right)$. Se habla entonces de T como del período de retorno y x_T es el evento de retorno. Cuanto mayor sea el período de retorno, T, mayor será el valor x_T, por encima del cual consideramos que se da un evento extremo.

Cuando X representa pérdidas, existe otra forma de definir un evento o riesgo extremo. Dada una probabilidad p, llamada probabilidad de excedencia, el Value at Risk (VaR) para p es el valor $VaRp$, tal que $P\,[X \geq VaR_p] = p$, es decir, $VaR_p = F^{-1}\,(1-p)$.

Por su parte, dada una probabilidad de excedencia p, se define el Tail Value at Risk (TVaR), también llamado Conditional Tail Expectation (CTE), como el promedio de las pérdidas mayores o iguales a VaR_p, es decir, $E[X \mid X > VaR_p]$. En finanzas, el TVaR también se conoce como *expected shortfall.*

3. DISTRIBUCIONES COMUNES EN LA MODELIZACIÓN DE EVENTOS EXTREMOS

En el contexto de la modelización de eventos extremos asociados a su frecuencia de ocurrencia, y relacionados con fenómenos medioambientales, vamos a considerar como modelos para las comparaciones algunas de las distribuciones más utilizadas en este ámbito (Stedinger *et al.*, 1993): concretamente, las distribuciones normal, log-normal triparamétrica, Gumbel, Generalized Extreme Value, Pearson Tipo 3, Log-Pearson Tipo 3 y Pareto Generalizada.

Por su parte, en el contexto de pérdidas por daños debidos a esos fenómenos, vamos a considerar las siguientes: log-normal triparamétrica, Pearson Tipo 3, Log-Pearson Tipo 3, Pareto Generalizada, Burr, Birnbaum-Saunders y Dagum; estas distribuciones aparecen, entre otras, como distribuciones habituales para datos de pérdidas (ver, por ejemplo, Dutang *et al.,* 2008).

4. LA DISTRIBUCIÓN PARETO POSITIVE STABLE

La distribución PPS (Sarabia y Prieto, 2009; Guillén *et al.*, 2011) puede ser definida por su función de supervivencia, dada por

$$S\,(x) = \exp\left\{-\lambda\left[\log\left(x/\sigma\right)\right]^{\nu}\right\}. \qquad (1)$$

Si $\nu = 1$, esta expresión define una ley de potencias, de manera que éstas son casos particulares de la distribución PPS. De la ecuación (1),

$$\log S(x) = -\lambda \left[\log (x/\sigma)\right]^{\nu},$$

que implica una relación potencial entre el logaritmo de la función de supervivencia y los valores de la variable que sólo es lineal cuando $\nu = 1$, es decir, cuando la PPS es realmente una distribución de Pareto o ley de potencia. Si tomamos logaritmos de nuevo en esa expresión,

$$\log \left[-\log S(x)\right] = \log \lambda + \nu \log \left[\log (x/\sigma)\right], \tag{2}$$

que puede ser interpretado como una relación lineal entre la función de supervivencia y los datos normalizados por el parámetro de escala σ, en una escala de doble log-log.

Todas las funciones habituales en el manejo de distribuciones de probabilidad, junto con la generación de valores aletatorios, la estimación de los parámetros por distintos métodos y el problema de la bondad del ajuste, en el contexto de la distribución PPS, se hayan implementados en la librería ParetoPositiveStable de R (Sáez-Castillo *et al.,* 2014).

5. LA PPS EN LA MODELIZACIÓN DE VALORES EXTREMOS RELATIVOS A FENÓMENOS NATURALES

En la Tabla 1 mostramos los valores del Criterio de Información de Akaike (AIC) y del estadístico Kolmogorov-Smirnov (KS), obtenidos del ajuste de las diferentes distribuciones de probabilidad consideradas en el presente trabajo, en datos relacionados con la ocurrencia de eventos medioambientales extremos: la distribución PPS es la más precisa en el ajuste por al menos uno de los dos criterios en tres de los cuatro ejemplos. Por su parte, la Figura 1 y la Figura 2 muestran el ajuste de los datos sobre daños ocasionados por huracanes (Sáez-Castillo *et al.,* 2011) y por inundaciones en Estados Unidos junto con la estimación de los períodos de retorno, VaR y TVaR para cada modelo ajustado.

6. CONCLUSIÓN

Dado que las distribuciones que se utilizan como modelos para la determinación de eventos extremos desempeñan un papel fundamental en la cuantificación del riesgo asociado a ellos, resulta fundamental contar con un amplio catálogo de tales distribuciones teóricas de entre la que escoger la más adecuada, según algún criterio estadístico de bondad de ajuste. Partiendo de que no existe la mejor distribución, válida en todas las aplicaciones y en todos los contextos, sí conviene tener presente modelos que muestren su capacidad de ajuste en distintos ámbitos. Creemos que los ejemplos mostrados en este trabajo, y otros que se han omitido por brevedad, ponen de manifiesto que la distribución PPS cumple con este requisito, ya que aporta ajustes adecuados en ámbitos tan diferentes como precipitaciones máximas, caudales máximos, velocidades máximas de viento o, en el contexto de las pérdidas, daños ocasionados por huracanes o inundaciones.

	Fort River		Fort Collins		Lisboa		Hardwich	
Dist.	AIC	KS	AIC	KS	AIC	KS	AIC	KS
PPS	1668.09	**0.06**	**1135.42**	**0.04**	247.52	**0.07**	22.74	0.09
NORM	1771.31	0.25	1170.95	0.12	**246.05**	0.08	30.33	0.13
LN	**1668.16**	0.06	1135.73	0.04	247.45	0.08		
GUMBEL	1725.89	0.20	1139.29	0.06	247.32	0.09	**19.13**	0.10
GEV	1697.20	0.22	1136.96	0.05	247.25	0.08	21.13	0.10
P3	1672.69	0.13	1135.62	0.05	247.44	0.08	21.94	**0.09**
LP3	1669.58	0.06	1136.15	0.04	247.45	0.08	21.32	0.09
GP	1677.71	0.64	1205.68		325.86		180.96	

Tabla 1. Ajuste de eventos medioambientales extremos en diferentes localizaciones: caudal anual máximo en Fort River (Estados Unidos), precipitación anual máxima en Fort Collins (Estados Unidos), velocidad anual máxima del viento en Lisboa y altura anual máxima del mar en Hardwich (Reino Unido). Se muestra el valor del AIC y el estadístico de Kolmogorov-Smirnoff (en cada caso, la distribución más precisa aparece en negrita).

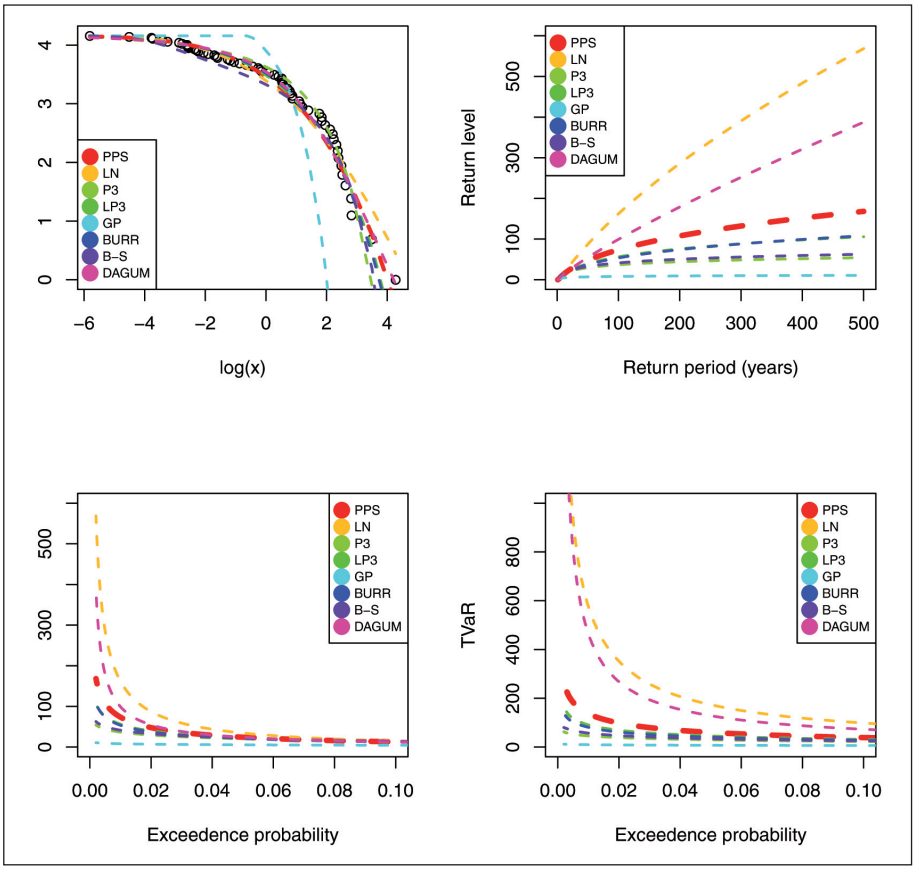


Figura 1. Ajuste del daño económico estimado (en billones de USD), ocasionado por huracanes en Estados Unidos anualmente desde 1926 a 1995, y medidas del riesgo asociadas a dichos ajustes.

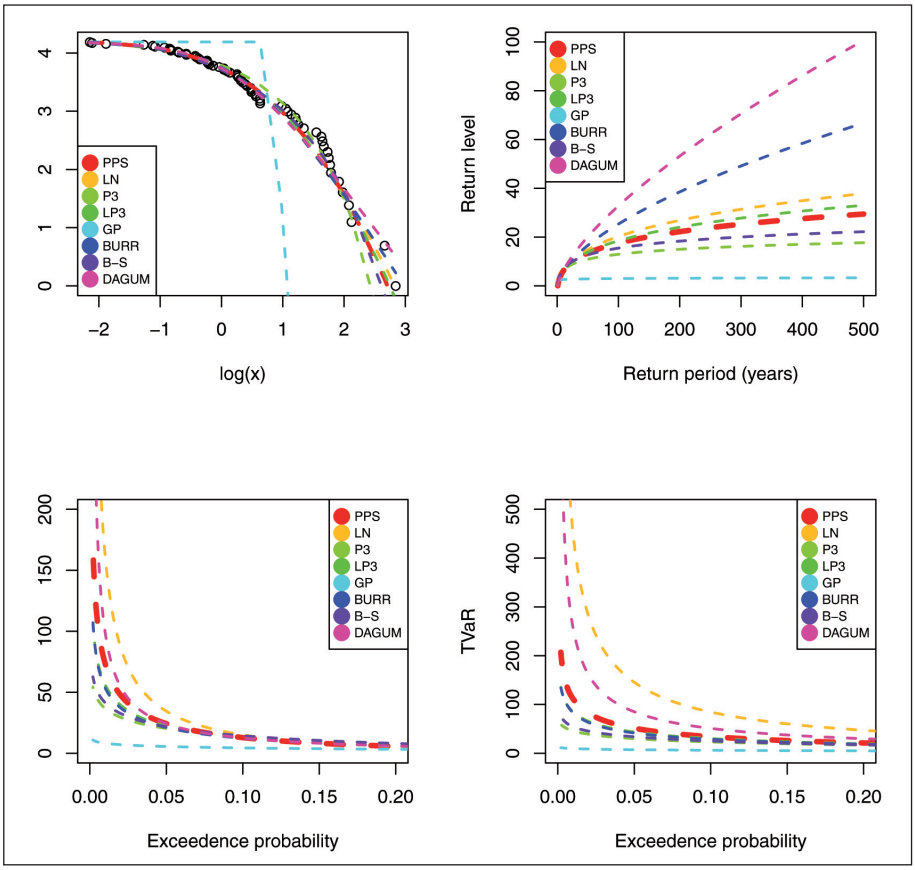


Figura 2: Ajuste del daño económico total (en billones de USD), ocasionado por inundaciones en Estados Unidos anualmente desde 1932 a 1997, y medidas del riesgo asociadas a tales ajustes

REFERENCIAS

Dutang, C., Goulet, V., y Pigeon, M. (2008). "Actuar: An R Package for Actuarial Science". *Journal of Statistical Software*, 25 (7), 1-37.

Guillen, M., Prieto, F., y Sarabia, J.M. (2011). "Modelling Losses and Locating the Tail with the Pareto Positive Stable Distribution". *Insurance: Mathematics and Economics*, 49 (3), 454-461.

Sáez-Castillo, A., Prieto, F., y Sarabia, J.M. (2014). *ParetoPosStable: Computing, Fitting and Validating the PPS Distribution*. ParetoPosStable R package version 1.0.3, <URL: http://CRAN.R-project.org/package=ParetoPosStable>.

Sáez-Castillo, A., Prieto, F., y Sarabia, J.M. (2011). "Modelización estadística mediante distribuciones PPS de pérdidas por daños debidas a riesgos extremos en función de algún regresor categórico". *Cuadernos de la Fundación Mapfre*, 171, 5-16.

Sarabia, J.M., y Prieto, F. (2009). "The Pareto-Positive Stable Distribution: A New Descriptive Model for City Size Data". *Physica A: Statistical Mechanics and its Applications*, 388 (19), 4179-4191.

Stedinger, J.R., Vogel, R.M., y Foufoula-Georgiou, E. (1993). "Frequency Analysis of Extreme Events". *Handbook of Hydrology*, edited by D. Maidment. New York: McGraw-Hill, Inc.

82. RISK AGGREGATION IN A CLASS OF DEPENDENT HEAVY-TAILED DISTRIBUTIONS

J.M. Sarabia
1 University of Cantabria, Department of Economics, Spain
sarabiaj@unican.es

F. Prieto
University of Cantabria, Department of Economics, Spain
faustino.prieto@unican.es

V. Jordá
University of Cantabria, Department of Economics, Spain
vanesa.jorda@unican.es

ABSTRACT

In this paper we obtain closed-form formulas for basic risk measures when we consider aggregated risks and when we work with a class of dependent risks with second kind beta marginals. We begin with the individual risk model, where we obtain the probability density function (PDF) and several risk measures, including VaR, TVaR and other tail measures. Then, we work with the collective risk model SN , and we obtain closed expressions for the PDF, assuming several primary distributions as Poisson, negative binomial and logarithmic distributions. Finally, and application with real data is included.

1. INTRODUCTION

The study of the distribution of the total claim amount is an ongoing topic of research in actuarial science. The aggregate amount of insured losses is given by a sum of N

individual claim amounts which are identically and independently distributed random variables. In the individual risk model, N is given so the subject of the analysis is the size of the losses. Elliptical distributions have been extensively used to model the amount of insurance loss. However, losses are asymmetrically distributed with a heavy right tail, features that cannot be represented by the normality assumption. Consequently, other parametric functional forms have been used to model aggregate risk, being most of them special cases of the Generalized Beta family (Guillén *et al.*, 2013).

On the other hand, the hypothesis of independence can be quite restrictive since there is well-known evidence of dependence between some kinds of risks (see Cosette *et al.*, 2002). Then, some studies have introduced dependence in the model using copulas and multivariate distributions (Albrecher *et al.*, 2011; Cossette *et al.*, 2008).

In the collective risk model, the number of losses is a random variable, which conventionally has been represented as a Poisson process. However, some portfolios may present over-dispersion and a high frequency of occurrence of zero claims (zero-inflated), which invalidates the Poisson distribution as a model for the frequency of claims. This fact has led the consideration of other discrete models such as the negative binomial.

In this work, we obtain closed-form formulas for basic risk measures of the aggregate loss. We begin the analysis with the individual model, assuming that the size of individual claims follows the beta distribution of second kind. We obtain the PDF, VaR and TVaR for this model. Thereafter, we consider the collective risk model, assuming several functional forms for the frequency of claims, including the Poisson, negative binomial and the Logarithmic distributions. We characterize the compounded distributions by their PDFs and a battery of risk measures is also given. Finally, we illustrate our methodology with an application to real data.

2. A CLASS OF DEPENDENT RISKS BASED ON THE SECOND KIND BETA DISTRIBUTION

In this section, we introduce a class of dependent risks with second kind beta marginals (see Guillén *et al.*, 2013). Let $Y_0, Y_2, \ldots, Y_m$ be mutually independent gamma random variables such that $Y_0 \sim \mathcal{G}(q_0)$ and $Y_i \sim \mathcal{G}(p_i)$, $i = 1, 2, \ldots, m$. The multivariate second kind beta distribution is defined as,

$$(X_1, X_2, \ldots, X_m)^\top = \left(\lambda_1 \frac{Y_1}{Y_0}, \lambda_2 \frac{Y_2}{Y_0}, \ldots, \lambda_m \frac{Y_m}{Y_0}\right)^\top, \tag{1}$$

where $q_0 > 0$ and $\lambda_i, p_i > 0$, $i = 1, 2, \ldots, m$. By construction $X_i \sim \mathcal{B}2(p_i, q_0, \lambda_i)$, $_i = 1, 2, \ldots, m$ with PDF $f(x_i; p, q, \lambda) \propto x_i^{p-1}(1 + x_i/\lambda)^{-(p+q)}$ if $x > 0$. The parameter q_0 introduces the dependence in the model.

3. THE INDIVIDUAL RISK MODEL

In this section, we obtain the distribution of the model of the individual risks according to the dependent model (1).

Theorem 3.1. Let $(X_1, X_2, \ldots, X_m)^{\mathrm{T}}$ be a multivariate second kind beta distribution defined as the stochastic representation (1). Then, if $\lambda_i = \lambda$ for all $i = 1, 2, \ldots, m$, the distribution of the individual risk $S_m = X_1 + \cdots + X_m$ is,

$$S_m \sim \mathcal{B}2(p_1 + p_2 + \cdots + p_m, q_0, \lambda). \tag{2}$$

Using (2) several risk measures can be obtained (see Guillén *et al.*, 2013). The VaR of (2) is,

$$VaR[S_m; \alpha, \lambda, \underline{p}, q_0] = \lambda \frac{B^{-1}(\alpha; p_1 + \cdots + p_m, q_0)}{1 - B^{-1}(\alpha; p_1 + \cdots + p_m, q_0)},$$

where $B^{-1}(\alpha; p, q)$ denotes the inverse of the incomplete ratio beta function and $0 < \alpha < 1$. Similar expression can be obtained for the TVaR and other tail risk measures.

4. COLLECTIVE RISK MODEL

In this section, we obtain the distribution of the compound collective model $S_N = X_1 + X_2 + \cdots + X_N$, with second kind beta as secondary distribution and taking several primary distributions.

4.1. Compound Poisson distribution

Let us consider the compound Poisson-Second kind beta model, where the primary distribution is Poisson with mean $\theta > 0$. Since the characteristic function of the second kind beta is (see Phillips, 1982),

$$\varphi_X(t; p, q) = \frac{\Gamma(p+q)}{\Gamma(q)} \Psi(p, 1-q, -it),$$

where $\Psi(z; a, c)$ is the confluent hypergeometric function of the second kind defined as,

$$\Psi(z; a, c) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1-t)^{c-a-1} dt,$$

the characteristic function of S_N is,

$$\varphi_{S_N}(t; \theta, p, q) = \exp\left\{ \theta \left(\frac{\Gamma(p+q)}{\Gamma(q)} \Psi(p, 1-q, -it) - 1 \right) \right\}. \tag{3}$$

The mean and variance of the compound Poisson-Second kind beta distribution are given by,

$$E(S_N) = \frac{\theta \lambda p}{q-1}, \quad q > 1, \tag{4}$$

$$var(S_N) = \frac{\theta \lambda^2 p(p+1)}{(q-1)^2(q-2)}, \quad q > 2. \tag{5}$$

The following theorem provides the PDF of the S_N distribution.

Theorem 4.1. If we assume a Poisson distribution with parameter θ as primary distribution and a second beta distribution with parameter (p, q, λ) as secondary

distribution, then the PDF of the random variable $S_N = X_I + \cdots + X_N$ for some selected p values is:

$$f_{S_N}(x;\theta,q,\lambda) = \frac{q(q+1)\theta e^{-\theta}}{\lambda} \frac{(x/\lambda)\ {}_2F_2\left[1+\frac{q}{2},\frac{3+q}{2};\frac{3}{2},2;\theta\left(\frac{x/\lambda}{1+x/\lambda}\right)^2\right]}{(1+x/\lambda)^{q+2}}, \tag{6}$$

if $x > 0$ *for* $p = 2$ *and*

$$f_{S_N}(x;\theta,q,\lambda) = \frac{q(q+1)(q+2)\theta e^{-\theta}}{2\lambda} \frac{(x/\lambda)^2\ {}_3F_3\left[1+\frac{q}{3},\frac{4+q}{3},\frac{5+q}{3};\frac{4}{3},\frac{5}{3},2;\theta\left(\frac{x/\lambda}{1+x/\lambda}\right)^3\right]}{(1+x/\lambda)^{q+3}}, \tag{7}$$

if $x > 0$ for $p = 2$ and $f_{S_N}(0) = e^{-\theta}$.

Similar results can be obtained taking a negative binomial distribution as primary distribution.

4.2. Compound Logarithmic distribution

Now we assume a logarithmic distribution with $\Pr(N = n) = \frac{\theta^n}{-\log(1-\theta)n}$, $n = 1, 2, \ldots$ as primary distribution. The characteristic function of S_N is,

$$\varphi_{S_N}(t;\theta,p,q) = \frac{\log\left\{1 - \theta\frac{\Gamma(p+q)}{\Gamma(q)}\Psi(p,1-q,-it)\right\}}{\log(1-\theta)}.$$

In the following theorem we have the PDF of the random variable S_N when $q = 1$.

Theorem 4.2. If we assume a logarithmic distribution with parameter θ as primary distribution and a second beta distribution with parameter *(p, q, λ)* as secondary distribution, then the PDF of the random variable S_N for $q = 1$ is given by,

$$f_{S_N}(x;\theta,p,\lambda) = \frac{-p\theta}{\lambda\log(1-\theta)} \frac{(x/\lambda)^{p-1}}{(1+x/\lambda)[(1+x/\lambda)^p - \theta(x/\lambda)^p]}, \quad x > 0. \tag{8}$$

5. EMPIRICAL APPLICATION TO EXTRAORDINARY RISKS

We considered historical data on losses (property damages) from extraordinary events ocurred in Spain (extraordinary floods; earthquakes; atypical cyclonic storms;

terrorism; riots and civil commotions; actions by armed or security forces in peaceful time) paid by the Insurance Compensation Consortium (Consorcio de Compensación de Seguros, CCS), a Spanish public corporate entity attached to the Ministry of Economy and Competitiveness. Data correspond to the total amounts (in euros updated Dec 31, 2013), paid by CCS for damage to properties, in each of the 50 spanish provinces and 2 spanish autonomous cities, in the years 2004, 2007, 2010 and 2012, published online by the Insurance Compensation Consortium (2014).

We fitted the Compound Logarithmic model (8) by the method of maximum likelihood. Table 1 shows a descriptive summary of the datasets. Table 2 shows the corresponding parameter estimates and values of Bayesian information criterion *(BIC)* statistic defined by (see Schwarz, 1978)

$$BIC = \log L - \frac{1}{2} d \log n; \tag{9}$$

where $\log L$ is the log-likelihood of the model evaluated at the maximum likelihood estimates, d is the number of parameters and n is the number of data.

Finally, as a graphical model validation, we used a rank-size plot on a log-log scale (see Prieto *et al.*, 2014). These plots appear in Figure 1 showing a good fit of the Compound Logarithmic distribution to data.

Dataset	n	Mean (10^6)	Std. Dev. (10^6)	Skewness	Kurtosis	Min.	Max. (10^6)
2004	52	2.73	6.28	3.68	17.18	2368	35.15
2007	52	6.11	14.27	3.92	19.11	4886	82.64
2010	52	9.05	12.87	1.96	5.68	49062	48.80
2012	52	5.23	15.70	3.67	15.14	586	74.55

Table 1. Main empirical characteristics of total losses by province datasets. (in euros updated Dec 31, 2013). Years: 2004, 2007, 2010 and 2012.

Dataset	$\hat{p}$	$\hat{\theta}$	$\hat{\lambda}$	BI C
2004	118.83	0.99797	120.02	-800.74
2007	131.87	0.99912	133.09	-835.47
2010	431.70	0.99235	431.95	-887.85
2012	33.831	0.99985	66.263	-796.65

Table 2. Parameter estimates from the Compound Logarithmic model to the total losses by province datasets by maximum likelihood. BIC statistics for the model considered. Years: 2004, 2007, 2010 and 2012.

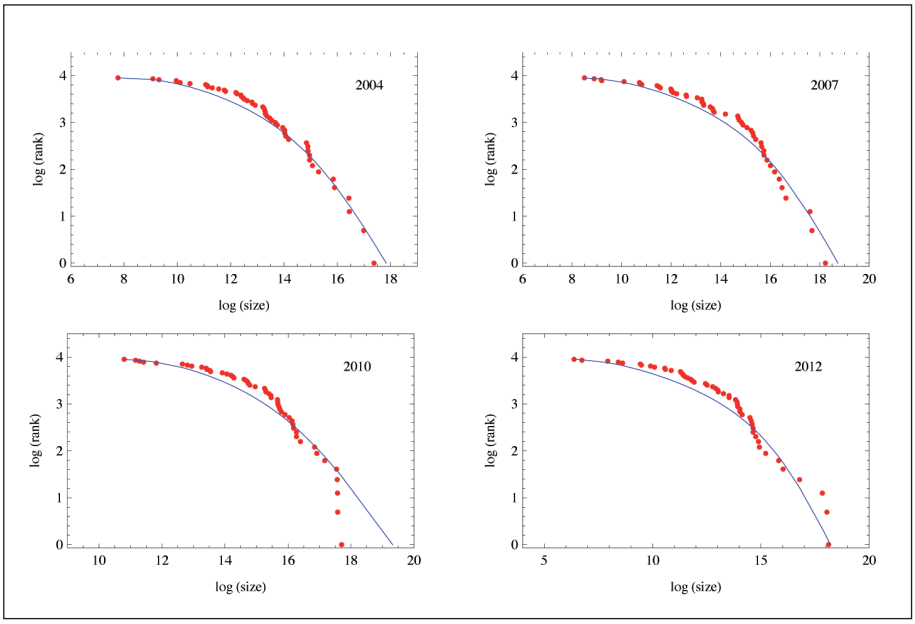


Figure 1. Rank-size plots of the complementary of the cumulative distribution function (cdf) multiplied by $n + 1$ (solid lines) of the Compound Logarithmic distribution and the observed data, on log-log scale. Years: 2004, 2007, 2010 and 2012.

ACKNOWLEDGMENTS

The authors thanks to Ministerio de Economía y Competitividad (projects ECO2010-5145 and ECO2013-48326-C2-2-P) for partial support of this work.

REFERENCES

Albrecher, H., Constantinescu, C., Loisel, S. (2011). "Explicit Ruin Formulas for Models with Dependence among Risks". *Insurance: Mathematics and Economics*, 48 (2), 265-270.

Consorcio de Compensación de Seguros (2014). *Estadística. Riesgos extraordinarios. Serie 1971-2013*. http://www.consorseguros.es/web/p_re

Cossette, H., Gaillardetz, P., Marceau, É., Rioux, J. (2002). "On two Dependent Individual Risk Models". *Insurance: Mathematics and Economics*, 30 (2), 153-166.

Cossette, H., Marceau, E., Marri, F. (2008). "On the Compound Poisson Risk Model with Dependence Based on a Generalized Farlie-Gumbel-Morgenstern Copula". *Insurance: Mathematics and Economics*, 43 (3), 444-455.

Guillén, M., Sarabia, J. M., Prieto, F. (2013). "Simple Risk Measure Calculations for Sums of Positive Random Variables". *Insurance: Mathematics and Economics*, 53 (1), 273-280.

Phillips, P.C.B. (1982). "The True Characteristic Function of the F Distribution". *Biometrika*, 69, 261-264.

Prieto, F., Sarabia, J.M., Sáez, A.J. (2014). "Modelling Major Failures in Power Grids in the Whole Range". *International Journal of Electrical Power & Energy Systems*, 54, 10-16.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics*, 6 (2), 461-464.

83. STATISTICAL TESTS FOR THE TAIL OF THE SEISMIC-MOMENT DISTRIBUTION OF GLOBAL SHALLOW EARTHQUAKES

I. Serra
1 Centre de Recerca Matemàtica, Barcelona, Spain
iserra@crm.cat

Á. Corral
Centre de Recerca Matemàtica, Barcelona, Spain
acorral@crm.cat

ABSTRACT

Gutenberg-Richter (GR) law is of fundamental importance in statistical seismology, stating that the earthquake-energy distribution is well described by a power law. We attempt to improve the fit given by the GR law using extensions of the power-law distribution and likelihood ratio tests, with new data including recent mega-earthquakes. A discussion of results will be presented.

1. INTRODUCTION

The size of earthquakes is measured either by the magnitude or the seismic moment. The relation between the (scalar) seismic moment, M, and the (moment) magnitude, m, of an earthquake is

$$m = \frac{2}{3}\left(\log_{10} M - 9.1\right) \tag{1}$$

when the seismic moment is measured in N·m, see Kagan (1997). The most well-known "stylized fact" in statistical seismology is the Gutenberg-Richter (GR) law. It

states that, for a given region, the magnitudes of earthquakes follow an exponential probability distribution, provided that the magnitude values are not too small to be poorly recorded. A simple change of variables shows that the seismic moment follows then a power-law, since M is an exponential function of magnitude. Therefore, the probability density function for a seismic moment bigger than a fixed positive threshold denoted by a is given by

$$f_{GR}(M) = \frac{\beta}{a}\left(\frac{a}{M}\right)^{1+\beta}; \tag{2}$$

where $\beta > 0$ is the shape parameter, also referred to as the exponent. Usually, β takes values close to 0.65, see Kagan (2010).

The power-law distribution has important physical implications, as it suggests an origin from a critical branching process or a self-organized-critical system, see Vere-Jones (1976), Bak (1996), and Bell *et al.* (2013). Nevertheless, it presents also some conceptual difficulties, due to the fact that the mean value $\langle M \rangle$ provided by the distribution turns out to be infinite. These elementary considerations imply that the GR law cannot be naively extended to arbitrarily large values of M, and one needs to introduce additional parameters to describe the tail of the distribution, coming presumably from finite-size effects. However, a big problem comes from the fact that the change from power law to a faster decay seems to take place at the highest values of M that have been observed, for which the statistics is very poor, see Zöller (2013).

From the statistical point of view, a solution to tame the power-law tail is to consider a weight given by the exponential function. The natural procedure consists in introduce it into the probability density function. This yields a left-truncated gamma (TrG) distribution with a negative shape parameter, whose probability density function is given by

$$f_{TrG}(M) = \frac{1}{\theta\Gamma(-\beta, a/\theta)}\left(\frac{\theta}{M}\right)^{1+\beta} e^{-M/\theta} \tag{3}$$

where $M > a, \beta > 0$, and $\theta > 0$, with θ the scale parameter of the exponential tail and with $\Gamma(\gamma, z) = \int_z^\infty x^{\gamma-1}e^{-x}dx$ the (upper) incomplete gamma function, for which $z > 0$ if $\gamma < 0$.

However, from a practical point of view it is useful to introduce the modulation given by the exponential function into the complementary cumulative distribution function. In the statistical seismology framework this is referred to as the tapered GR (TGR) law and its probability density function is

$$f_{TGR}(M) = \left[\frac{\beta}{a}\left(\frac{a}{M}\right)^{1+\beta} + \frac{1}{\theta}\left(\frac{a}{M}\right)^{\beta}\right] e^{-(M-a)/\theta}; \tag{4}$$

where $M > a$, $\beta > 0$ and $\theta > 0$. *TrG* and *TGR* distributions were considered by Kagan (2002), who showed that both of them fulfilled the requirements necessary for any extension of the GR law.

Different results about the most appropriated law have been found depending on the time period and the statistical tools employed. Kagan (2002), with global data from 1977 to 1999, concludes that the TGR distribution gives a slightly better fit than the TrG, but this period does not include any event above magnitude 8.5. Main and coworkers have re-examined the problem using more recent data, see Main *et al*. (2008) and Bell *et al*. (2013), concluding that TGR gives a significantly better fit than GR for the period before the 2004 Sumatra-Andaman mega-earthquake, and arriving at the opposite conclusion for the period after this event.

As the previous authors, we analyze the Centroid Moment Tensor (CMT) global earthquake catalog, but for the period between 1 January 1977 and 31 October 2013. We restrict to shallow events (depth < 70 km) and, in order to avoid incompleteness, to magnitude $m > 5.75$ or, equivalently, $M > 10^{3/2 \cdot 5.75+9.1} = 5.3 \cdot 10^{17}$ Nm.

2. METHODOLOGICAL APPROACH

The objective is to get concluding results for the comparison of the three models given by GR, TrG and TGR distributions. Therefore, we are going to use the information obtained from the likelihood function evaluated at its maximum for each set of parameters.

Given a data set, $\{x_i\} i = 1, \cdots, n$, maximizing the likelihood under a wrong model is equivalent to find the best model as the closest one to the true distribution in the

sense of the Kullback-Leibler distance, see Pawitan (2001). In this sense, the best model is determined by likelihood inference. In fact, only the likelihood ratio (LR) between two models is relevant since this is invariant with respect to one-to-one transformation of the observed data, see Pawitan (2001). Therefore, the main statistic for comparing two models, M_1 and M_2, is

$$l_R = l_{M_2} - l_{M_1}$$

for $l_{M_i} = \sum_{i=1}^{n} \log f_{M_i}(x_i; \hat{\Theta}_i)$ where f_{Mi} denotes the probability density function of M_i and $\hat{\Theta}_i$ denotes the MLE for the set of parameters Θ_i, for every i = 1, 2. In order to compare the models, it is necessary characterizing the distribution of l_R.

Let p_1 and p_1 be the number of free parameters in the model M_1 and M_2, respectively. If the models are nested, then the probability distribution of the statistic $2l_R$ is approximately a chi-squared distribution with degrees of freedom equal to $p_2 - p_1 \geq 0$. This is the best test in the sense that it has the quickest convergence to the asymptotic distribution, see McCullagh and Cox (1986), and it is denoted by LR test. Note that the null and alternative hypothesis correspond to accept model M_1 or M_2, respectively.

If M_1 is nested within the boundary of the parameters of M_2, then we have to correct this distribution since it becomes a mixture of chi-squared distributions, see Castillo and Puig (1999). The GR is nested both in the TGR and in the TrG, as taking $\theta \to \infty$ in any of them leads to the power-law distribution. Moreover, the GR is in the boundary for both; then the GR model will be rejected if

$$2l_R > 2.71 \tag{5}$$

with a level of risk equal to 0.05, where $l_R = l_{TGR} - l_{GR}$ or $l_R = l_{TrG} - l_{GR}$, further details in Castillo and Puig (1999).

If the models are not nested, there is not a best methodology as above, but there are good solutions, for instance Akaike information criterium (AIC) and Bayesian information criterium (BIC) are designed for comparing the fit for several models. If the number of free parameters is the same in both models, then analyzing the comparison of the statistic AIC or BIC for two models is equivalent to analyze $2l_R$.

The method proposed for the non-nested case is the approach explained by Clauset *et al*. (2009) and denoted here as weakLR test. The TGR and TrG are not nested but have the same number of parameters, then the weakLR test is appropriated. In this case the critical values are depending on the sample size, n, and we accept that there exists a significative difference between the models if

$$|l_R| > 1.96\sqrt{n}s \tag{6}$$

where s denotes the standard deviation of the set

$$\{\log f_{TGR}(x_i, \hat{\beta}, \hat{\theta}) - \log f_{TrG}(x_i, \hat{\beta}, \hat{\theta})\}$$

for $i = 1, \cdots, n$, $l_R = l_{TrG} - l_{TGR}$ and the level of risk equal to 0.05. Further details in Clauset *et al*. (2009). In this case, note that the alternative hypothesis correspond to accept that the difference between the fit provided by the models is significative.

3. CONCLUDING RESULTS

Figure 1 shows the results obtained from the analysis of the global data set. The analysis consists in comparing the fit for every model analyzing the data from 1 January 1977 until the year indexed in x-axis. The first picture contains the critical region and the corresponding statistic, $2l_R$, for the LR test applied to compare GR and TGR models. Analogously, the second picture contains the results for GR and TrG and the last picture contains critical region and the corresponding statistic, l_R, for the weakLR test applied to compare TrG and TGR models. The main conclusion is that GR law can be rejected as a best model from 1980, see first and second picture. Moreover, the fit of the models TGR and TrG does not exhibit significative difference, but after the big event in 2004 the TrG presents a better fit, see the last picture. Further, in this case the GR is close to be not rejected in front of TGR, see the first picture. Then to cover the big events it is better to consider TrG. But this conclusion can not be extended, since the analysis of the fit is done for different tectonic zones and for all data and the results are not equal.

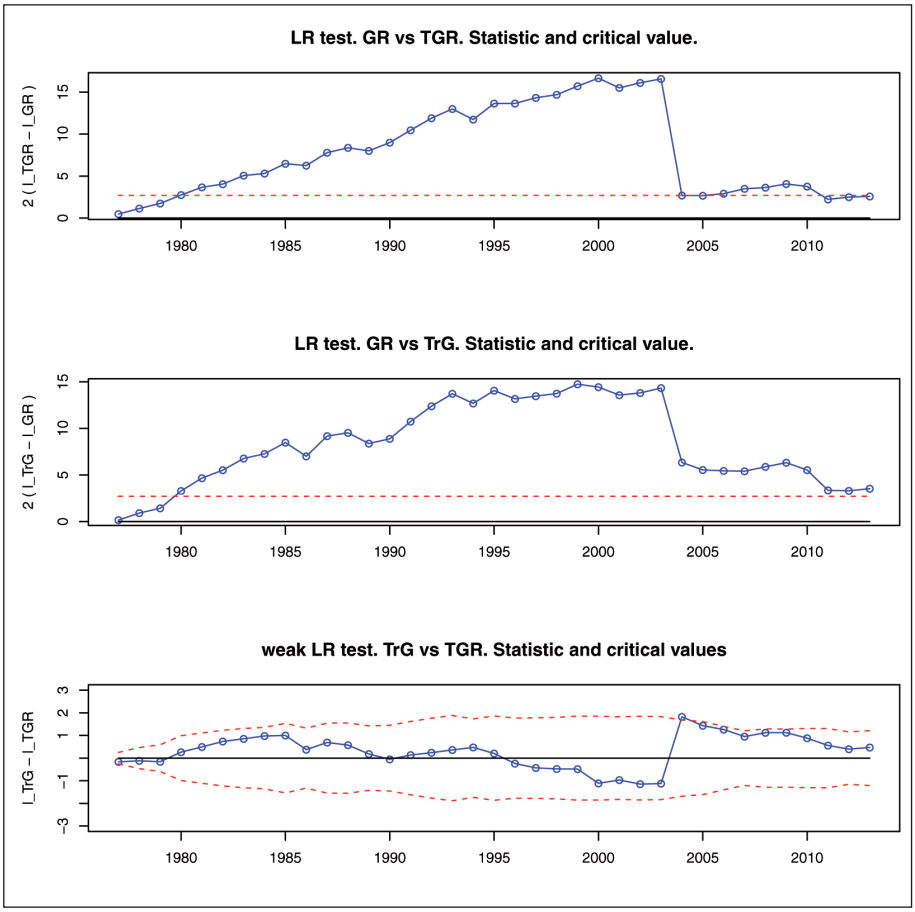


Figure 1. The dotted lines show the critical value for the corresponding test. The points are the statistic l_R or $2l_R$, that is labeled in y-axis, for each year from 1977 until 2013.

ACKNOWLEDGMENTS

We are grateful to M. Naylor for his feedback. Research expenses were founded by projects FIS2012-31324 and MTM2012-31118 from Spanish MINECO, 2014SGR-1307 from AGAUR, and the Collaborative Mathematics Project from La Caixa Foundation.

REFERENCES

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*. Copernicus, New York.

Bell, A.F., Naylor, M., and Main, I.G. (2013). "Convergence of the Frequencysize Distribution of Global Earthquakes". *Geophys. Res. Lett.*, 40, 2585-2589.

Castillo, J. del, and Puig, P. (1999). "The Best Test of Exponentiality against Singly Truncated Normal Alternatives". *Journal of the American Statistical Association*, 94, 529-532.

Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009). *Power-Law Distributions in Empirical Data*. SIAM Rev., 51, 661-703.

Kagan, Y.Y. (1997). "Seismic Moment-Frequency Relation for Shallow Earthquakes: Regional Comparison". *Journal of Geophysical Research*, 102, 2835-2852.

Kagan, Y.Y. (2002). "Seismic Moment Distribution Revisited: I. Statistical Results". *Geophys. J. Int.*, 148, 520-541.

Kagan, Y. Y. (2010). "Earthquake Size Distribution: Power-Law with Exponent $\beta \equiv 1/2$?". *Tectonophys.*, 490, 103-114.

Main, I.G., Li, L., McCloskey, J., and Naylor, M. (2008). "Effect of the Sumatran Mega-Earthquake on the Global Magnitude Cut-off and Event Rate". *Nature Geosci.*, 1, 142.

McCullagh, P., and Cox, D.R. (1986). "Invariants and Likelihood Ratio Statistics". *Ann. Statist.*, 14, 1419-1430.

Pawitan, Y (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Vere-Jones, D. (1976). "A Branching Model for Crack Propagation". *Pure Appl. Geophys.*, 114, 711-725.

Zöller, G. (2013). "Convergence of the Frequency-Magnitude Distribution of Global Earthquakes: Maybe in 200 Years". *Geophys. Res. Lett.*, 40, 3873-3877.

84. A SHARED-PARAMETER JOINT MODEL FOR PROSTATE CANCER RISK AND PSA LONGITUDINAL PROFILES

C. Serrat
Universitat Politècnica de Catalunya-BarcelonaTECH, Dept. of Applied Mathematics I, Spain
carles.serrat@upc.edu

M. Rué, H. Perpiñán and A. Forte
Universitat de Lleida, Dept. of Basic Medical Sciences, Spain
montse.rue@cmb.udl.cat

C. Armero
Universitat de València, Dept. of Statistics and OR, Spain,
{carmen.armero, hector.perpinan, anabel.forte}@uv.es

X. Piulachs
Universitat de Barcelona, Dept. of Econometrics, Spain,
xavierpiulachs@hotmail.com

Á. Páez
Hospital Universitario de Fuenlabrada, Urology Dept., Spain,
alvaro.paez@salud.madrid.org

G. Gómez
Universitat Politècnica de Catalunya-BarcelonaTECH, Dept. of Statistics and OR, Spain,
lupe.gomez@upc.edu

ABSTRACT

The paper describes the use of frequentist and Bayesian shared-parameter joint models of longitudinal measurements of prostate specific antigen (PSA) and the risk of prostate cancer (PCa). The motivating dataset corresponds to the screening arm of the Spanish branch of the European Randomized Screening for Prostate Cancer (ERSPC) study. The results show that PSA is highly associated with the risk of being diagnosed with PCa and that there is an age-varying effect of PSA on PCa risk. Both the frequentist and Bayesian paradigms produced very close parameter estimates and subsequent 95% confidence and credibility intervals. Dynamic estimations of disease-free probabilities obtained using Bayesian inference highlight the potential of joint models to guide personalized risk-based screening strategies.

1. INTRODUCTION

Joint models for longitudinal and time-to-event data are increasingly used to assess relationships between serial measurements of one or several markers and time to an event of interest. Joint models were introduced during the 90s (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997) and since then have been applied to a great variety of studies in epidemiological and biomedical areas. In turn, these studies have fed a wide methodological research on the subject, with models focused on event times, longitudinal patterns, or both (Neuhaus *et al.* (2009) and Tsiatis and Davidian (2004) are excellent reviews up to date).

In a setting of dependent longitudinal and time-to-event data, shared-parameter models consider that the longitudinal and survival processes depend jointly on a common set of random effects. Given the random effects, the two processes are assumed independent. Shared-parameter models allow one to quantify the effect of the underlying longitudinal outcome on the risk of an event, and obtain individualized time-dynamic predictions. Recently, Rizopoulos has made a great contribution facilitating the use of the joint modeling methodology, first by means of an overview of the theory and applications of joint modeling (Rizopoulos, 2012) and secondly by developing the `JM` (Rizopoulos, 2010) and `JMbayes` (Rizopoulos, 2013) `R` packages for the frequentist and Bayesian shared-effects' approaches, respectively.

	Number of visits per subject				
	1	2	3	≥ 4	Overall
Sample size (%)	573 (23.7)	1499 (62.1)	293 (12.1)	50 (2.1)	2415 (100.0)
PSA descriptive					
Mean	2.40	1.40	1.92	5.46	1.78
StDev	5.67	1.67	2.25	2.79	3.20
Min	0.00	0.00	0.00	0.50	0.00
1st Q	0.60	0.60	0.60	3.79	0.63
Median	1.10	1.00	1.10	4.69	1.06
3rd Q	2.20	1.67	2.44	6.28	1.95
Max	68.90	42.30	20.36	15.53	68.90
PCa diagnosed (%)	51 (44.0)	50 (43.1)	13 (11.2)	2 (1.7)	116 (100.0)

Table 1. Distribution of the subjects in the screening arm of the Spanish ERSPC study, descriptive statistics of the PSA measurements and distribution of the PCa diagnosed cases, stratified by the number of visits and overall.

Proust-Lima and Taylor (2009) assessed the prognostic value of longitudinal PSA measurements as a marker of PCa progression, in patients treated with radiotherapy. Joint modeling is an interesting approach that allows to illustrate how classical and Bayesian statistics can be combined to reach a complex goal, taking advantage of the strengths of both paradigms. The goal of the paper is to show, from both approaches, how shared-parameter joint models can be used to incorporate past and current PSA levels into a predictive model of PCa, in a screening setting.

2. ERSPC STUDY AND DATA DESCRIPTION

Our motivating dataset consists of the screening arm of the Spanish subset of the European Randomized Screening for Prostate Cancer (ERSPC) study (Luján *et al.*, 2012). The ERSPC study is a multicenter trial initiated in 1991 in the Netherlands and Belgium, with six more European countries joining during the 90s. The objective of the ERSPC was to evaluate whether PCa screening with PSA reduces PCa specific mortality in asymptomatic men. The Spanish subset recruited 4278 individuals between February 1996 and June 1999, aged 45 to 71 years, of whom 2415 were assigned to the screening arm and the remaining to the control arm. Those screened were given a blood test to measure PSA levels. Table 1 summarizes the description of the data.

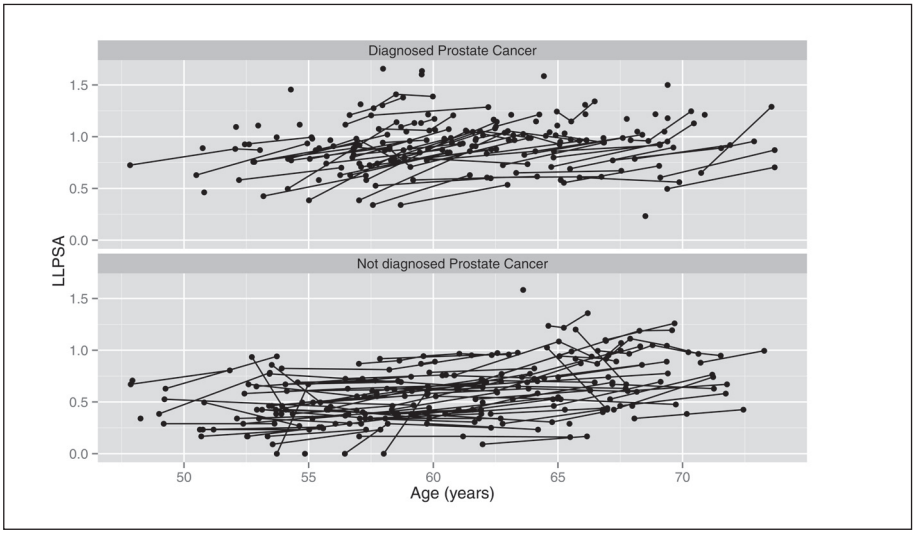


Figure 1. *LLPSA* profiles versus time (age in years) for men who developed PCa during follow-up (top) and for 116 randomly selected men without a PCa diagnosis (bottom), participants in the screening arm of the Spanish ERSPC study.

We defined the time-to-event, T, as the time elapsed from the protocol screening start (age 45) to diagnosis of prostate cancer. We estimated that more than 90% of patients were prostate cancer free at age 80 years. In particular 116 men (4.8%) were diagnosed with PCa (median time, over age 45, equal to 17.32 years) and the remaining 2,299 men (95.2%) contributed to the study as right-censored data.

3. MODEL PROPOSAL

For the i-th subject, $i = 1, \ldots, n$, denote by $m_i(t)$ the true $LLPSA$ value at time t (i.e. at age $45 + t$) and by $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s \leq t\}$ the whole longitudinal history of the true marker levels up to time t. The observable data for the i-th subject at the specific occasions t_{ij} at which measurements were taken, consists of the observed longitudinal $LLPSA$ profile $\{y_{ij} = LLPSA_{ij} = \log(1 + \log(1 + PSA_{ij})), j = 1, \ldots, n_i\}$. Let $y_i(t)$ be the hypothetical observed value of $LLPSA$ at time t which is assumed to deviate from the true value $m_i(t)$ by a certain error measurement $\varepsilon_i(t)$ in such a way that $y_i(t) = m_i(t) + \varepsilon_i(t)$ holds.

The joint model that we propose binds a linear mixed-effects model for the longitudinal part and a relative risk survival model for the time to PCa diagnosis, which incorporates the true historical subject profile of the longitudinal process (Rizopoulos, 2012). The survival submodel is based on the Weibull distribution due to its flexibility for representing different types of risks. To the best of our knowledge, this is the first time that a straightforward shared-parameter joint model of PSA and age at PCa diagnosis is fitted, in a screening setting, with the objective of exploring if longitudinal PSA measurements may orientate a personalized risk-based screening. The joint model is specified as

$$\begin{cases} (y_{ij} \mid \boldsymbol{\theta}_i) = m_i(t_{ij}) + \varepsilon_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij} \\ h_i(t|\mathcal{M}_i(t), w_{1i}, \boldsymbol{\theta}_i) = h_0(t \mid \lambda)\exp\{\gamma_0 + \gamma_1 w_{1i} + \alpha m_i(t)\}, \end{cases} \quad (1)$$

where $h_i(t) = \lim_{dt\to 0} \Pr(t \le T_i^* < t + dt \mid T_i^* \ge t)/dt$ denotes the hazard function, T^* denotes the true event time, and

$$\boldsymbol{\theta}_i = (\beta_0, \beta_1, b_{i0}, b_{i1}, \sigma_{b_0}, \sigma_{b_1}, \rho_{b_0 b_1}, \sigma, \lambda, \gamma_0, \gamma_1, \alpha)^{\mathsf{T}}$$

is the vector containing all parameters, random effects and hyperparameters associated with the individual i. For the longitudinal submodel, $(\beta_0, \beta_1)^{\mathrm{T}}$ and $(b_{i0}, b_{i1})^{\mathrm{T}}$ are the fixed and the random effects for the intercept and the slope term, respectively, and ε_{ij} the error term for the i-th subject at the j-th measurement with $(\varepsilon ij \mid \sigma) \sim \mathcal{N}(0, \sigma^2)$; furthermore we assume that both random effects given σ_{b0}, σ_{b1} and their correlation coefficient, $\rho_{b0\,b1}$, follow a bivariate normal distribution with a mean of zero and an unstructured co variance matrix. For the survival submodel, the baseline hazard is given by $h_0(t \mid \lambda) = \lambda\, t^{\lambda-1}$, where λ is the shape parameter; $w_1 = a_1 \times y_1$ is a covariate which describes the interaction between the age, a_1, and the LLP SA measurement, y_1, at first visit, respectively; $(\gamma_0, \gamma_1)^{\mathrm{T}}$ are the regression coefficients for the baseline hazard and the w_1 interaction. Finally, α stands for the association parameter assessing the relationship between the longitudinal and the survival submodels.

Since preliminary analysis for joint model (1) showed a large correlation between b_{i0} and b_{i1} we discarded the slope random effect and restrict posterior analyses, both frequentist and Bayesian, interpretations and conclusions to the following model

$$h_i(t|\mathcal{M}_i(t), w_{1i}, \boldsymbol{\theta}_i) = h_0(t \mid \lambda)\exp\{\gamma_0 + \gamma_1 w_{1i} + \alpha(\beta_0 + b_{i0} + \beta_1 t)\}, \quad (2)$$

where $(b_{i0} \mid \sigma_{b_0}) \sim \mathcal{N}(0, \sigma_{b_0}^2)$.

4. ESTIMATION PROCEDURE, RESULTS AND INTERPRETATION

In the frequentist approach, the joint model expression in (2) was fitted through the `jointModel` function, implemented in the `R` package `JM` (Table 2). The estimates reflect a strong positive association between PSA values and risk of PCa, and allows estimation of the impact of different increments of PSA at specific PSA values.

Parameters	Value	95% Conf. Int.
Longitudinal Submodel		
β_0	0.369	(0.351, 0.388)
β_1	0.014	(0.013, 0.015)
σ_2	0.011	(0.011, 0.012)
σ^2_{b0}	0.050	(0.044, 0.057)
Survival Submode		
γ_0	-13.706	(-15.299, -12.113)
γ_1	-0.068	(-0.107, -0.028
λ	1.887	(1.459, 2.315)
Association		
α	7.207	(6.041, 8.372)

Table 2. Joint model estimates for frequentist analysis of PCa diagnosis and longitudinal PSA values.

Similar results are obtained when estimating the full vector of uncertainties in the model, $\boldsymbol{\theta} = \cup_{i=1}^{n} \boldsymbol{\theta}_i$, from a Bayesian perspective, by using non-informative distributions as a priors and the common prior distribution given by

$$\pi(\boldsymbol{\theta}) = \pi(\lambda)\,\pi(\gamma_0)\,\pi(\gamma_1)\,\pi(\alpha)\;\pi(\beta_0)\,\pi(\beta_1)\,\pi(\mathbf{b}_0 \mid \sigma_{b_0})\,\pi(\sigma_{b_0})\,\pi(\sigma).$$

Computations were carried out using the statistical software `WinBUGS` (Lunn *et al.*, 2000) and `R`-package `JMbayes`.

The proposed methodology allows to dynamically estimate disease-free probabilities. Figure 2 is an illustration of these updated survival probabilities based on the observed longitudinal subject-specific marker profile available up to date.

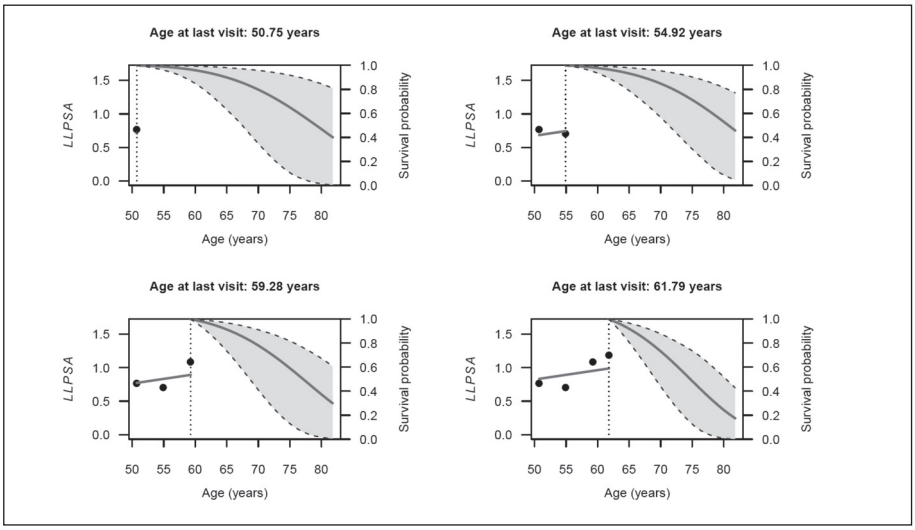


Figure 2. *LLPSA* longitudinal trajectories and dynamic PCa free survival probabilities (with 95% pointwise credible intervals at each time point) for a patient after their first to fourth successive PSA measurements.

5. DISCUSSION

In this study, using data from the Spanish arm of the ERSPC study, we have estimated a joint model of longitudinal measurements of PSA and time to diagnosis of PCa. Our model overcomes the limitations of simpler statistical tools because it accounts for a) the effect of PSA as an endogenous time-dependent covariate measured with error and b) the non-random dropout that results when an individual is diagnosed with PCa. In particular, we used both the frequentist and the Bayesian approaches, and built a shared-parameter joint model that combined some baseline covariates and longitudinal PSA values to predict PCa incidence. The resulting model is consistent with the literature and clinical knowledge. This work can be considered a first step that should be followed by a more comprehensive modeling process accounting for additional predictive factors measured over time, e.g. prostate volume or previous biopsy status (Roobol *et al.*, 2012).

ACKNOWLEDGMENTS

This paper has been partially supported by research grants MTM2012-38067-C02-01 and MTM2013-42323-P from the Spanish Ministry of Economy and Competitiveness and 2014 SGR 464 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya.

REFERENCES

Faucett, C.L., and Thomas, D.C. (1996). "Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach". *Statistics in Medicine*, 15, 1663-1685.

Lujan, M., Paez, A., Berenguer, A., and Rodriguez, J.A. (2012). "Mortality Due to Prostate Cancer in the Spanish Arm of the European Randomized Study of Screening for Prostate Cancer (ERSPC). Results after a 15-Year Follow-Up". *Actas Urológicas Españolas*, 36, 403-409.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility". *Statistics and Computing*, 10, 325-337.

Neuhaus, A., Augustin, T., Heumann, C., and Daume, M. (2009). "A Review on Joint Models in Biometrical Research". *Journal of Statistical Theory and Practice*, 3, 855-868.

Proust-Lima, C., and Taylor, J.M. (2009). "Development and Validation of a Dynamic Prognostic Tool for Prostate Cancer Recurrence Using Repeated Measures of Posttreatment PSA: A Joint Modeling Approach". *Biostatistics*, 10, 535-549.

Rizopoulos, D. (2010). JM: "An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data". *Journal of Statistical Software*, 35, 1-33.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. CRC Press, Boca Ratón, FL.

Rizopoulos, D. (2013). *JMbayes: Shared Parameter Models for the Joint Modeling of Longitudinal and Time-to-Event Data Using JAGS, WinBUGS, or OpenBUGS*. R package 0.4-1 http://cran.r-project.org/package=JMbayes/.

Roobol, M.J., Van Vugt, H.A., Loeb, S., Zhu, X., Bul, M., Bangma, C.H., Van Leenders, A.G., Steyerberg, E.W., and Schroder, F.H. (2012). "Prediction of Prostate Cancer Risk: The Role of Prostate Volume and Digital Rectal Examination in the ERSPC Risk Calculators". *European Urology*, 61, 577-583.

Tsiatis, A.A., and Davidian, M. (2004). "Joint Modeling of Longitudinal and Time-to-Event Data: An Overview". *Statistica Sinica*, 14, 809-834.

Wulfsohn, M.S., and Tsiatis, A.A. (1997). "A Joint Model for Survival and Longitudinal Data Measured with Error". *Biometrics*, 53, 330-339.

85. THEORY OF ECONOMIC SYSTEMS AS BASE FOR STRUCTURING SPACE OF RISK FACTORS

R. Kachalov
Central Economics and Mathematics Institute, Moscow, Russia
kachalov@cemi.rssi.ru

Y. Sleptsova
Dubna International University for Nature, Society and Man, Dubna, Russia
julia_sleptsova@mail.ru

ABSTRACT

Article is devoted to the concept of structuring space economic risk factors based on view of economic processes as the transformation processes of economic systems. Three directions of system economics application for fulfillment of the strategic tasks have been defined.

1. INTRODUCTION

To improve the quality of managing a socioeconomic system one must react efficiently to external and internal impacts and detect changes within relevant economic risk factors (ERF) in due time. The theory of economic systems (Kleiner, 2011, 2012) provides adequate tools for that. This research relies on this theory and suggests both an instrumental base for unified structuring risk space of any level socioeconomic system and a systematic method of detecting and identifying ERF. There existing four core economic systems: object, environmental, project and process (Kleiner, 2013), let's consider the object one.

Here are some examples of this system:

1. Production cluster (Porter, 1990) implies a located geographically group of affiliated companies, specialized suppliers, service providers, financial organizations and companies in related business sectors and also organizations related to their activity.

2. Chain outlet is a company with a module structure, a new organizational pattern in global economy (Castels, 1996-1998), its particular functions being distributed among its subsidiaries.

3. Financial industrial group (FIG) which can incorporate self-consistent (industrial, trade, transport, credit) companies in terms of their business and legal activities. FIGs are created by virtue of the decree of the President (Russian Federation Presidential Decree from December 5 1993. No 2096) through integrating financial and material resources of companies of different profile.

Taking into account the time-space index, one can single out subsystems of the same four types within the economic system concerned. The subsystem of the object type shall be a business or a group of businesses comprising the socioeconomic system (SES) concerned. Project type subsystems realized at this particular period of time by the subsystems of object type and their structural elements shall also be part of the system. Both subsystems realize certain internal value transfer processes within the initial system and in interaction with external economic systems, which results in various and (or) evolutional changes in their condition. Finally, environmental subsystem type can be mentioned. It deals with active legislation at working places or places where end products are sold, service is provided and other economic activity is implemented.

Based on the theory of economic systems and ERF space structuring caused by it, a new approach can be found.

This approach enables us to solve the following tasks:

1. Tasks to realize strategic solutions. A forecast of socioeconomic condition on all four subsystems (object, environmental, process, project) is made on each

unspecified stage of strategic solution realization at Tn. Many economic risk factors comply with this condition.

2. Tasks of mid-term planning. In difference from the first group these tasks do not depend on making solutions within the system and can undermine the performance of socioeconomic systems if not properly dealt with, for instance, changing current legislation related to the operation of a socioeconomic system. The management of a company or a corporate group should react to these changes, a new condition of other subsystems of SES possibly arising. A variety of ERF which might be different from the variety of ERF at the previous moment should be brought in conformity with the new condition of SES. This particular type of tasks shall be considered in more detail here.

3. Tasks aimed at building up risk-based internal control at businesses. The current condition of SES is considered when solving these tasks. Afterwards, subsystems of SES in dynamics are compared to find out whether new projects and deals related to a business as a whole have come up over the last period of time, whether the implementation of the new legislation has taken place over the last period, how the changes in institutional environment have affected the projects and processes of the SES concerned. If no changes on the socioeconomic level, a checkup of relevance of detected risk factors goes to a lower level where ERF first appeared.

2. ERF IDENTIFICATIONS WAYS

This report deals with both detecting ERFs and working out anti-risk controlling actions in solving tasks of mid-term planning for cases related to changes in current legislation (changes of the component of SES environment).

Let's discuss some existing methods of ERF identification. Later on it will be shown what applies the theory of economic systems. Nowadays a number of methods to detect risk factors can be applied. In the operational theory of risk management there are universal methods used by audit and insurance companies and special ones that allow particular risk assessment while doing the

research (Kachalov, 2012). An example of this method shall be a combination of two ways of detecting risk factors in describing a risk situation (Sleptsova, Kachalov, 2014).

The first way, shall we refer to it as a "top down" way, uses existing standards of risk management, among them the classification of risk factors on the pattern suggested in FERMA standards (http://www.ferma.eu/risk-management/standards/risk-management- standard). The advantage of this way is that the main ERF will be analyzed considering the place where they arise for SES, but the detailed analysis of this method is insufficient.

The second way – shall we refer to it as "down top". In terms of organizational structure, any business consists of a number of subdivisions. Each subdivision and its employees participate in some business processes. Some ERFs correspond to each business process. The shortcoming of this second way is that essential ERFs dealing with business activities on the whole are hard to be formalized within particular business processes inside the organizational structure and so can be lost.

3. THE SPECIFIC CASE

On the first stage let's consider how the other SES components will change when the environment changes. For instance, the tax legislation has changed and individual tax payers have caught an opportunity to pay income tax collectively. For that purpose a contract about creating a consolidated group of tax payers is necessary and some formalities are to be complied with, i.e. additional processes arise. The organizational structure of a business can change or can remain unchanged depending on the decision the management will take. New funds for investment can appear if a business follows the new pattern of tax calculation and tax payment. In classifying factors of economic risks based on SES components four groups of risks can be singled out: object risks, process risks, projects risks and environmental risks. ERFs detected at T_0 are assigned to SES condition at T_0. Should the environment change (Eenvironment$_0$ changes to Environment$_1$), a managerial decision to change processes (Process$_0$ changes to Process$_1$) will

be taken. In our case a business will transfer all the tax registers and supporting documents from a local area to the head office where the consolidated group of tax payers is registered. The environment change in this case will arise not only due to putting new legislation into operation but also in case of considerable payments to the local budget at T_0. At the moment T_1 the business will not transfer part of tax to the local or regional budget which might change its relations with representatives from local authorities. Therefore a detailed situational analysis is to show whether new ERFs will or will not arise. Anyway ERF with correct application of new legislation is to arise. Document transfer processes risk factors along with data exchange risk factors between a local business and a head office can be referred to new ERF processes. Decisions dealing with changes in organizational structure and whether new ERF can be identified there are also to be taken into account. In increasing funds allocated for investment, should this decision be made, one should consider what project ERFs will be detected or will lose its relevance due to the termination of a project stage before schedule. Working out and endorsement of regulations and procedures can be applied as anti-risk controlling action to ERF processes (see Figure 1).

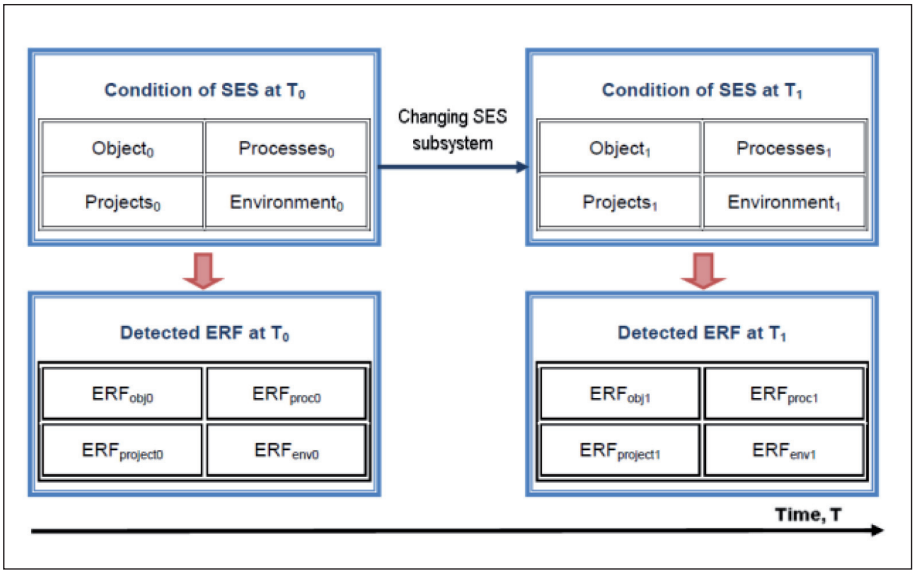


Figure 1. Application of the theory of economic systems to ERF.

5. CONCLUSION

In conclusion we note finally that if one makes a forecast to the conditions of SES subsystem while planning stages of strategic solutions realization, the application of the theory of economic systems will allow us to identify ERFs more efficiently and increase the quality of SES management.

In case of mid-term planning the theory of economic systems will enable us to make use of the systematic approach to detecting new ERFs and choose anti-risk controlling actions.

In controlling the fullness and consistency of financial accounting, meeting legislative norms the theory of economic systems allows us to analyze the changes in the four subsystems urgently and assess the condition of internal control system in the previous period.

Thus, applying the theory of economic systems, grouping ERFs into four subsystems enables us to improve the quality of SES management: detect new ERFs more efficiently at a strategic planning stage, actualize the system of anti-risk controlling actions while in operation and optimize the procedure of internal control.

ACKNOWLEDGEMENTS

This work has been supported by Russian Humanitarian Science Foundation (project Nº 14-02-00333).

REFERENCES

Castells, M. (1996-1998). *The Information Age: Economy, Society and Culture*. Vol. I-III. Cambridge, MA; Oxford, UK: Blackwell.

Decree of the President of Russian Federation (1993). "About Creation Financial and Industrial Groups", of Dec. 5, no. 2096.

Kachalov, R. (2012). *Managing Economic Risk: Theoretical Background and Applications*. St. Petersburg. 288 p.

Kleiner, G. (2011). "A New Theory of Economic Systems and its Applications". *Herald of the Russian Academy of sciences*, v.81, no. 5, 516-532.

Kleiner, G., Lankin V., and Arutunova D. (2012). "The Process Approach to Economic Systems Interaction Modeling". *ATINER's Conference Paper Series*. ECO2012-0275. Athens Institute for Education and Research (ATINER). Athens. http://www.atiner.gr/papers/ECO2012-0275.pdf

Porter, M.E. (1990). *Competitive Advantage of Nations*. New York: Free Press (2nd. ed., New York: Free Press, 1998. 896 p. ISBN 978-0-684-84147-2).

Sleptsova, Y., Kachalov, R. (2014). "Quantitative Rating of Economic Risk Level in Business Operation". *Nauchno-technicheskie vedomosti of Saint Petersburg State Polytechnical University*. Economical sciences, 3 (197): 164-170.

86. COST RISK ANALYSIS METHODS FOR DEFENCE ACQUISITION PROJECTS

A. Sokri
DRDC CORA, Canada,
abderrahmane.sokri@drdc-rddc.gc.ca

A. Ghanmi
DRDC CORA, Canada,
ahmed.ghanmi@drdc-rddc.gc.ca

ABSTRACT

Cost risk analysis is a fundamental process in defence acquisition projects. It allows analysts to examine the impact of individual risks on overall project cost and to estimate the contingency budget amount needed for a desired level of confidence. Different methods for conducting cost risk analysis in defence acquisition projects have been developed in the literature. This paper provides a comprehensive historical review of the predominant research developments in risk analysis from a cost perspective. The paper discusses the underlying assumptions of the different cost risk analysis approaches, their methodologies, and how they can be used to estimate the contingency allowances.

1. INTRODUCTION

Military assets are generally appraised at the acquisition stage on the basis of their total life cycle costs. Cost estimates for major defence acquisitions are usually accompanied by a formal risk analysis. Indeed, point estimates (median or average expected cost) of these acquisitions have frequently underestimated the true cost by a wide margin. The reasons for this include changes in the specifications or

requirements as the program progresses, optimistic estimates of advances in technology, funding instability, price/cost escalation, and schedule slippage. Although a formal analysis of cost risk will not necessarily reduce the risk inherent in a program, it will help program managers understand the nature of the risks involved, and quantify and display the uncertainty associated with cost estimates. The result is a more realistic assessment of the funding required for a project and of the likelihood of exceeding reported point estimates.

Cost risk analysis is a fundamental process in project management. It provides a means to examine the impact of individual risks on the overall project cost, to determine the likelihood of finishing the project on budget, and to estimate the requisite contingency reserve needed for a desired level of certainty about achieving the cost plus reserve. Cost risk analysis can also identify the most important risks to project cost so risk mitigation can be effective. The analysis of cost risk can be conducted in the project's conceptual development phase as soon as there is a notional budget, and should be updated periodically throughout project execution as the estimate is refined and additional risks are identified and quantified.

The objective of this paper is to provide a comprehensive historical overview of the evaluation of risk assessment methodologies and to highlight recent research developments in risk estimation techniques. It also offers guidelines for applying appropriate statistical and operational research techniques in cost risk assessment.

2. EVOLUTION OF COST RISK ASSESSMENT METHODOLOGIES IN DEFENCE ACQUISITION PROJECTS

The existing literature on project cost risk analysis has primarily focused on the estimating techniques. Many researchers have contributed to the development of an overall framework. The range of its fields of application includes the environment, finance, project management, and cost estimating (Arena *et al.*, 2006). The military use of cost risk analysis can be traced back to the early 1950s after the end of World War II. Three main periods can be distinguished in this history: the multidisciplinary approach period (1950-1959), the econometric techniques period (1960-1980), and the stochastic simulation period (1990 to present).

2.1. Multidisciplinary approach period (1950-1959)

In the 1950s, scientists were mainly using a multidisciplinary approach including descriptive statistics, probability theory, and mathematical tools to analyze the equipment cost overrun factors (Marshall and Meckling, 1959). Over this decade, the most influential event was the introduction of the Program Evaluation Review Technique commonly abbreviated PERT. This statistical tool was developed by the United States Navy in 1957 to analyze the involved tasks in completing large and complex projects. Uncertainty is caused by scheduling a given project without having the details of all the activities. It can be reduced by better understanding of dependencies and improved scheduling of activities (Morris, 1994).

2.2. Econometric techniques period (1960-1989)

In the period 1960-1989, cost overruns on many high-visibility public projects started to shed light on the importance of cost risk analysis (Morris, 1994). Probabilistic risk analysis started to be recognized as an approved part of policy analysis (National Research Council, 1983). During this period of transition, the American Society for Risk Analysis has expanded to become an international organization. The advancement of statistical techniques and the growth of computing power led to more-refined techniques of cost risk analysis.

One of the econometric techniques is the cost growth technique. This technique applies econometric techniques to the costs of a set of historical projects to derive a cost growth factor. The difference between the revised and the estimated costs and the width of the confidence intervals are used as measures of the overall cost risk (Bolten *et al*., 2008). Dienemann (1966), for example, treated the future cost of a system as a random variable. This random variable is characterized by its probability density function. Integrating this variable between its lowest possible value and some value of interest gives the probability that the final cost will be at or below that value.

2.3. Stochastic simulation period (1990 to present)

In the period 1990 to the present day, personal computers and software packages became widespread and their power continued to increase rapidly. This made the

use of stochastic simulation methods for proactive decision making in project evaluation possible. Simulation uses mathematical models, probability distributions, and random sampling to draw instances of results.

There are two main stochastic simulation methods for cost risk analysis (Hulett, 2012): Cost driver and risk driver methods. The cost driver method uses a cost breakdown structure of a project and represents each cost line item by a probability distribution indicating the variability in the cost estimate. The risk driver method addresses the weaknesses of the cost driver method by using a risk breakdown structure instead of a cost breakdown structure to perform cost risk analysis. The cost driver method has three main limitations: (1) A cost element may be influenced by several different risks and a risk may impact multiple cost elements. (2) Correlation between cost elements exists in most projects. Using the cost driver method requires the specification of correlation coefficients to represent this phenomenon. These correlation coefficients are particularly difficult to obtain. (3) The method focuses on the impact of risks rather than the risks themselves. It particularly ignores the probability that a risk project event may or may not occur.

A growing body of literature has begun to recognize simulation and convolution as approved methods for proactive decision making in military project evaluation (Arena *et al.*, 2006; US Air Force, 2007; NASA, 2008; NATO, 2012; Maybury, 2014). Sokri and Solomon (2014), for example, used a probabilistic risk assessment to portray the major factors of uncertainty and estimate the Next Generation Fighter Capability (NGFC) project cost contingency. The authors showed that the single-figure expected value approach yields a higher predictable value of cost contingency than Monte Carlo simulation.

This period also saw the widespread development by the International Cost Estimating and Analysis Association (ICEAA) of new standards in terminology and application of analysis techniques. The Scenario-Based Method is one of the techniques that ICEAA published in its official professional journal (Garvey *et al.*, 2012).

In parallel to all the developments mentioned above, three simplified approaches have usually been used: The percentage addition method, the expected monetary value, and the qualitative methods. The percentage addition method determines

contingency as a percentage addition on the base estimate of a project. This subjective value is sometimes at the 100th percentile guaranteeing enough funds for any upward random variation in project cost (Sokri and Solomon, 2014). The expected monetary value estimates contingency as probability times consequence. Kaplan and Garrick (1981) found that the expected monetary value view reduces risk to the expected value of damage. They concluded that this viewpoint alone is not able to express the idea of risk. Qualitative methods use subjective judgments and ordinal scaling techniques to prioritize risks. Risk probability and risk impact are generally described in non-numerical terms such as low, moderate, and high (Hulett, 2012).

3. MAJOR LIMITATIONS

The key decision point in selecting a cost risk method over another is the availability of reliable and relevant data that will be used to assess cost risk. A multidisciplinary approach including a qualitative cost risk analysis would be appropriate early in the project life cycle when there is no or little information about the project. An econometric approach would be appropriate early in the project life cycle when some historical data and limited project information are available. Use of a simulation approach would be recommended when sufficient data on risk drivers or a detailed design specification is available (NASA, 2008).

The multidisciplinary approach evaluates each risk to determine how likely it would affect the project cost if it occurs. A major critical consideration of this approach is that it does not take into account the potential correlations among the cost risk drivers. Such a correlation may exacerbate or dilute the aggregated incremental cost. The qualitative part of this approach is based on subjective judgments to prioritize risks. This qualitative analysis should be thought of as a quick and temporary evaluation that should be reviewed as soon as new information is available.

A major limitation of the econometric approach is that it requires a valid relationship between cost and the different risk drivers. The accurate and homogenous historical data required for a formal econometric analysis is not always accessible. The lack of a significant similarity between historic systems and the new one is another major disadvantage of this method.

The level of similarity is generally based on subjective judgement made by analysts.

While both the stochastic simulation and convolution methods can be used to generate a new aggregate probability distribution of the project cost estimate and what is known as an S-curve, they present different benefits and limitations. Convolution involves solving analytical formulas to derive closed-form solutions. Three main approaches are used to combine distributions analytically: manipulation of integrals, moment generating functions, and characteristic functions. The moment generating function, for example, does not always exist. Sometimes its integral does not converge. Simulation can combine all input distributions to generate a possible distribution of the outcome, but it is still unable of determining other aspects of cost risk such as the identification of the most likely risk tolerance level and the determination of the residual risk (United States Government Accountability Office, 2009).

4. CONCLUSION

Cost risk analysis is a fundamental process in defence acquisition projects. It examines the impact of individual risks on the overall project cost and estimates the contingency amount needed for a desired level of certainty. Different methods for analyzing cost risk in defence acquisition projects have been developed. Their military use can be traced back to the early 1950s. These techniques can be separated along historical lines into three main periods: the multidisciplinary approach period (1950-1959), the econometric techniques period (1960-1989), and the stochastic simulation period (1990 to present).

In addition to a comprehensive historical review of these techniques, this paper discusses their underlying assumptions, and describes their limitations. The paper mainly showed that the most difficult step in assessing a risky project is to combine the distributions of different risk factors. There are two major approaches to aggregating risk distributions: stochastic simulation and analytical approach. The simulation approach is very flexible and provides detailed quantitative analysis. The analytical method is very difficult and not practical for most uses and users. Further efforts should be undertaken to address other aspects in cost risk analysis. Any methodology

used for the derivation of the estimate should be able to determine a specific risk tolerance and evaluate the remaining risk after mitigation.

REFERENCES

Arena, M.V., Younossi, O., Galway, L.A., Fox, B., Graser, J.C., Sollinger, J.M. *et al*. (2006). *Impossible Certainty Cost Risk Analysis for Air Force Systems*. RAND Corporation, MG-415. CA, USA.

Bolten, J.G., Leonard, R.S., Arena, M.V. Younossi, O., and Sollinger, J.M. (2008). *Sources of Weapon System Cost Growth*. RAND Corporation, MG-670. CA, USA.

Dienemann, P.F. (1966). *Estimating Cost Uncertainty Using Monte Carlo Techniques*. RAND Corporation, RM-4854-PR. CA, USA.

Garvey, P.R., Flynn, B., Braxton, P., and Lee, R. (2012). "Enhanced Scenario-Based Method for Cost Risk Analysis: Theory, Application, and Implementation". *Journal of Cost Analysis and Parametrics*, 5(2), 98-142.

Hulett, D. (2012). *Integrated Cost-Schedule Risk Analysis*. Gower. Burlington, USA.

Kaplan, S., and Garrick, B.J. (1981). "On the Quantitative Definition of Risk". *Risk Analysis*, 1(1), 11-27.

Marshall, A.W., and Meckling, W.H. (1959). *Predictability of the Costs, Time, and Success of Development*. RAND Corporation, P-1821. CA, USA.

Maybury, D. (2014). *MEOSAR Cost Escalation Risk*. DRDC-RDDC Scientific Letter 2014-L14-0224-1455(E). Ottawa, Canada.

Morris, P.W.G. (1994). *The Management of Projects.* Thomas Telford. London, UK.

NASA (2008). *Cost Estimating Handbook* (CEH). Washington DC, USA.

National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process.* National Academy Press. Washington D.C., USA.

NATO (2012). *NATO Independent Cost Estimating and the Role of Life Cycle Cost Analysis in Managing the Defence Enterprise*. NATO RTO Technical Report TR-SAS-076.

Sokri, A., and Solomon, B. (2014). *Cost Risk Analysis and Contingency for the NGFC*. DRDC CORA Technical Memorandum TM 2013-224. Ottawa, Canada.

United States Government Accountability Office (2009). *Cost Estimation and Assessment Guide*. GAO-09-3SP. Washington D.C., USA.

US Air Force (2007). *Cost Risk and Uncertainty Analysis Handbook*. MA, USA.

87. ON MULTIVARIATE EXTENSIONS OF THE INCREASING CONVEX ORDER TO COMPARE RISKS

M. A. Sordo
University of Cádiz, Department of Statistics and Operation Research, Spain,
mangel.sordo@uca.es

ABSTRACT

In this work, we propose a generalization of the increasing convex order to the multivariate setting to compare portfolios of risks. This order accounts both for the marginal effects and the dependence structures of the vectors and is closely related to some multivariate risk measures recently introduced in the literature. We derive the main properties of this new stochastic order.

1. INTRODUCTION

In insurance, an important reason for quantifying losses in the tail of distributions is to compare risks and, for this purpose, different risk measures and stochastic orders can be used. One of the most popular risk measures is the conditional tail expectation. Let X be a random variable with distribution function F and corresponding quantile function $F^{-1}(p) = \inf\{x : F(x) \geq p\}$, for $p \in (0, 1)$. The conditional tail expectation (CTE) of X at quantile p is defined by

$$\mathbf{CTE}_p(X) = E\left[X \,|X > F^{-1}(p)\right], \; p \in (0,1).$$

The CTE is a coherent risk measure that represents the expected risk given that the total risk exceeds its p-quantile. For continuous distributions, the CTE coincides with the tail value-at-risk given by

$$\mathbf{TVaR}_p(X) = \frac{1}{1-p}\int_p^1 F^{-1}(t)\,dt, p \in (0,1), \tag{1}$$

(see Chapter 2 in Denuit *et al.*, 2005). The tail value-at-risk is closely related to the increasing convex order (also named the stop-loss order), which is defined as follows.

Definition 1.1. Given two random variables X and Y with distribution functions F and G, X is said to be smaller than Y in the increasing convex order (denoted by $X \leq_{icx} Y$) if $E[u(X)] \leq E[u(Y)]$ for all increasing convex function u.

Roughly speaking, if $X \leq_{icx} Y$, the risk X is simultaneously smaller and less variable than Y. It is well-known that $X \leq_{icx} Y$ if and only if

$$E\left[(X-t)^+\right] \leq E\left[(Y-t)^+\right], \text{ for all } t,$$

where $x^+ = \max\{x, 0\}$. In insurance, the function $\pi(t) = E(X-t)^+$ represents the net premium of a stop-loss reinsurance contract, therefore the increasing convex order is a natural way of comparing the riskiness of two univariate random variables. Many other properties and applications of this order can be found in the books by Müller and Stoyan (2002), Denuit *et al.* (2005) and Shaked and Shanthikumar (2007). The increasing convex order is characterized in terms of the tail value-at-risk as follows

$$X \leq_{icx} Y \text{ if and only if } \mathbf{TVaR}_p(X) \leq \mathbf{TVaR}_p(Y) \text{ for all } p \in (0,1) \tag{2}$$

(see, for example, Lemma 2.1 in Sordo and Ramos, 2007, which is stated as Theorem 4.A.3 in Shaked and Shanthikumar, 2007). Therefore, for continuous random variables, we have

$$X \leq_{icx} Y \text{ if and only if } \mathbf{CTE}_p(X) \leq \mathbf{CTE}_p(Y) \text{ for all } p \in (0,1) \tag{3}$$

a result that interprets the comparison of CTEs in terms of evaluation functionals of the form $E[u(X)]$ and $E[u(Y)]$, where u is any increasing convex function.

In the literature, a number of authors, including Tibiletti (1993), Landsman and Valdez (2003), Hürlimann (2004) and Cousin and Di Bernardino (2014) have suggested extensions of the univariate CTE to the multivariate context. In general,

these extensions provide different translations of the notion of "tail distribution" at level p, with $p \in (0, 1)$, to the multivariate setting (see also Sordo, Suárez-Llorens and Bello, 2015, for this topic). In this work we suggest a multivariate generalization of the increasing convex order that preserves some properties of the univariate increasing convex order, including a natural extension of (3) that involves the multivariate CTE introduced by Cousin and Di Bernardino (2014). Given two random vectors $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$, the new stochastic order accounts both for the marginal effects and the dependence structures of the vectors. In particular, for comonotonic random vectors, this order reduces to the univariate increasing convex order of the marginals. Among other properties, it will be shown that if $\mathbf{X}$ and $\mathbf{Y}$ have the same conditionally increasing structure (which is a notion of positive dependence) the increasing convex order among the marginals X_i and Y_i, for $i = 1, \ldots, n$, implies the multivariate order among the vectors $\mathbf{X}$ and $\mathbf{Y}$. As a consequence, we extend an increasing-convex order preserving property for the multivariate CTE from the case of Archimedean copulas (as provided by Hürlimann, 2014) to the case of conditional increasing copulas. [Remark: this work is still in progress at the time of writing this report].

REFERENCES

Cousin, A., and Di Bernardino, E. (2014). "On Multivariate Extensions of Conditional-Tail Expectation". *Insurance: Mathematics and Economics* 55, 272-282.

Denuit, M., Dhaene, J., Goovaerts, M.J., and Kaas, R. (2005). *Actuarial Theory for Dependent Risks*. John Wiley&Sons.

Hürlimann, W. (2004). "Multivariate Fréchet Copulas and Conditional Value-at-Risk". *International Journal of Mathematics and Mathematical Sciences* 7, 345-364.

Hürlimann, W. (2014). "On some Properties of two Vector-Vauled VaR and CTE Multivariate Risk Measures for Archimedean Copulas". *ASTIN Bulletin* 44, 613-633.

Landsman, Z.N., and Valdez, E.A. (2003). "Tail Conditional Expectations for Elliptical Distributions". *North American Actuarial Journal* 7, 55-71.

Müller, A., and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, New York.

Shaked, M., and Shanthikumar, J.G. (2007). *Stochastic orders*. Series: Springer Series in Statistics, Springer

Tibiletti, L. (1993). "On a New Notion of Multidimensional Quantile". *Metron International Journal of Statistics* 51, 77-83.

Sordo, M.A, Suárez-Llorens, A., and Bello, A. (2015). "Comparisons of Conditional Distributions in Portfolios of Dependent Risks". *Insurance: Mathematics and Economics* 61, 62-69.

88. FRACTAL BASED CANCER MODELLING

M. Stehlík
JKU Linz/Universidad Técnica Federico Santa María,
IFAS/Departamento de Matemática, Austria/Chile
(corresponding) mlnstehlik@gmail.com

P. Hermann
JKU Linz, IFAS, Austria,
philipp.hermann@aon.at

ABSTRACT

Fractal hypothesis is both challenging and technical issue of mammary cancer. Here we discuss on statistical distributions of fractal dimensions for both mammary cancer and masthopaty.

1. INTRODUCTION

When we consider fractal based cancer diagnostic, many times a statistical procedure to assess the fractal dimension is needed. We shall look for some analytical tools for discrimination between cancer and healthy ranges of fractal dimensions of tissues (see Baish and Jain, 2000; Mrkvicka and Mattfeldt, 2011). The algebraic and topologic properties of cancer growth are available via appropriate set structure, e.g. bornology (see Paseka, Solovyov and Stehlík, 2015, 2014, or topology; see Stehlík, 2015). Here we illustrate some issues on discrimination between mammary cancer and masthopatic tissues, which is follow-up of study Hermann *et al*. (2014).

2. SIMPLE DISCRIMINATION BETWEEN MASTHOPATHY AND MAMMARY CANCER BASED ON THE BOX-COUNTING DIMENSION

We consider boxplots in Figure 1 in order to have a graphical comparison between the two groups. Therein, one can see that the box-counting dimensions seem to be on average lower for mammary cancer tissue in addition to some outliers in the lower boundaries.

If we will follow the simple concept that higher dimension is more risky, the issue is that we will arrive with this dataset to some sort of contradiction. The problem is that the median of the box-counting dimension is 1.5972. When we make a simple clustering based on ordering the box-counting dimension and decide to tell that more risky tissue have a box-counting dimension bigger than the median and non-risky tissue is below, then we only classified 135 of mamca and 60 of mastho below. Recall that 199 observations contain the characteristic masthopathy and 192 observations mammary cancer.

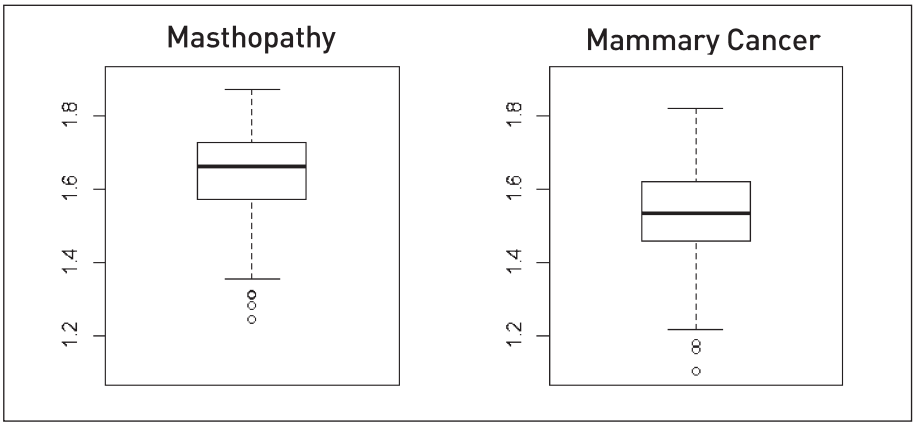


Figure 1: Boxplot of the groups masthopathy (left) and mammary cancer (right).

Even using the arithmetic mean of 1.587391 decreases the number of classified cancer tissues to 128 for mamca and 56 for masto. Based on this simple example we can conclude that we need a more sophisticated procedure based on the box-counting dimension to discriminate between the two groups and we should take more detailed

characteristics of the tissue into account. In extremal case there is no possibility to develop automatic clustering based on box-counting dimension, which can avoid histological expert examination.

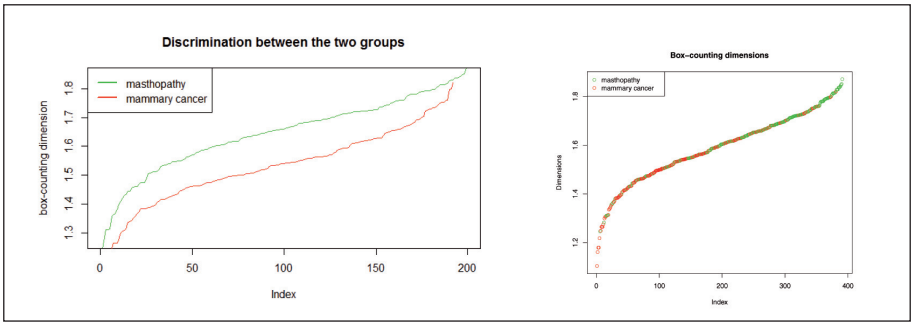


Figure 2 (left). Plot of the dimensions discriminated between the groups masthopathy and mammary cancer. Figure 3 (right). Plot of the dimensions.

Figure 2 indicates that to force usage only of one box-counting dimension establishes inverse problems, which is ill posed. Loosely saying we need a continuous dimension spectrum, e.g. multi-fractal dimension spectra. It has already been used in breast cancer discrimination by George and Kamal (2007). A multifractal system is a generalization of a fractal system in which a single exponent (the fractal dimension) is not enough to describe its dynamics; instead, a continuous spectrum of exponents (the so-called singularity spectrum) is needed. This relates also to Tweedie exponential dispersion models, which, as a special case, contains both normal and gamma distributions. This is further justification for these two simple distributional families: in the case of our empirical data we have found a strong deviation from normality for masthopathy, and therefore we used gamma and Weibull distribution. In contrast to that mammary cancer data has been adjusted by outliers which yields data following normal distribution.

Figure 3 gives a good overview of the distribution of the observations of the different groups. It is visible that mammary cancer tissue have on average higher dimensions (red) compared to masthopathological tissue (green). These conclusions were already recognizable due to the comparison of the mean and the interquartile-distance as well as within the boxplots.

To use the fact, that the sum of squared independent standard normal distributed random variables follows Chi-Squared distribution, it is essential to prove first this conditions. For this reason it is necessary to show that the differences between the curves are standard normal distributed. Therefore, Shapiro-Wilk-test was used and results in a rejection of the null-hypothesis of standard normal distribution. Hence, there is another possibility to proof the condition in order that usage of Chi-Squared distribution is allowed. We have to show, that the squared differences are Chi-Square distributed with one degree of freedom.

Computing the sum of the squared differences delivers a value of 5.38. A Chi-Square-test was accomplished to test whether the two groups are equal or different. The p-value of the distribution function of the Chi-Squared distribution with 199 degrees of freedom is approximately one. This value is another proof that the two groups are different. Furthermore, we made a standardization (by reducing of the mean and dividing by the standard error) of the previously calculated differences. The distribution function at the sum of standardized squared differences of 198 and 199 degrees of freedom is 0.49331. Hence, this p-value does not yield enough support to reject the null hypothesis of differences between the groups. However, we believe that the distribution can be gamma with the real valued shape parameter. We simulated in order to maximize p-values with changes in shape and scale parameters of Gamma distribution. By reducing the shape parameter and in contrast to that increasing the rate, which reduces the scale parameter, a convergence to higher p-values is obvious.

We can see that we will find a rather good fit for specific values of the parameters. These values were tested by the version of Kolmogorov-Smirnov in R. Thereby, the shape parameter of 0.415 and a scale parameter of 1/0.485 delivered an accurate p-value of 0.09956. The test with a shape parameter of 0.42 delivered an even better p-value of 0.1272. Therefore, it can be assumed that the standardized differences are Gamma distributed with a shape parameter lying in between the range [0.415, 0.42] and the scale parameter is about 2.06 (1/0.485). Moreover, it is quite useful to compare the standardized differences with generated random variables of a gamma distribution, with a shape parameter of 0.415 and a scale parameter of 2.062. This comparison can be seen in Figure 4.

shape	scale	p-value
0.45491680	1 / 0.45729929	8.354e-05
0.48	1 / 0.4573	3.7e-12
0.44	1 / 0.4673	1.65e-11
0.44	1 / 0.4673	5.55e-10
0.42	1 / 0.48	2.22e-16
0.43	1 / 0.47	1.403e-10
0.425	1 / 0.48	1e-9
0.42	1 / 0.48	0.004855
0.425	1 / 0.48	0.04488
0.43	1 / 0.48	0.009656
0.42	1 / 0.485	0.1272
0.415	1 / 0.485	0.09

Table 1. Simulation of p-values with given shape and scale parameter.

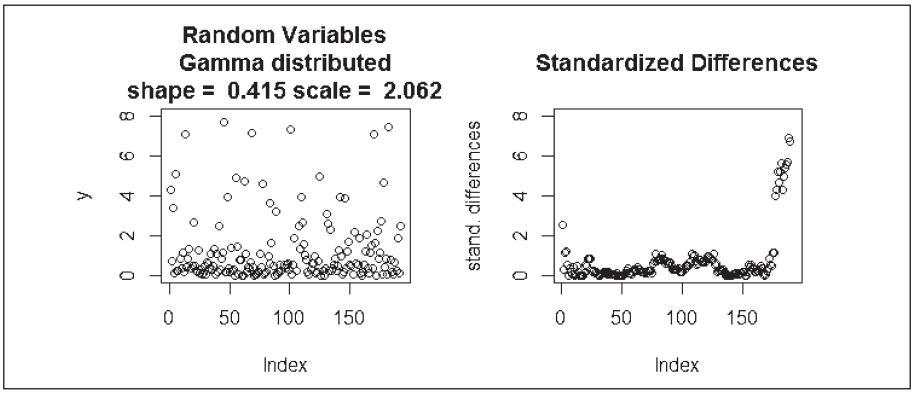


Figure 4. Comparison of standardized differences with random variables of a gamma distribution.

2.1. Testing for distributions of the groups

Shapiro-Wilk tests deliver p-value for mammary cancer tissue (0.04519) which is ten times the one resulting from mastophatic tissue. However, this value is still too small (for a significance level of 95%) to state that the box-counting dimension of mammary cancer tissue is normal distributed. QQ-Plots in Figure 5 are another proof, that masthopathy is not normally distributed, but normal distribution of the dimensions of mammary cancer should not be rejected without further analysis. Indeed, the lower quantile differs significantly from the comparative line.

Therefore, outlier detection for mammary cancer data with usage of box-plot rule has been performed. Four candidates for outliers from the lower end of the data were obtained and removed in order to yield useful information on the distributional behaviour of mammary cancer tissue. The according p-value has improved significantly up to 0.5716 and therefore, it can be assumed that the modified data is normal distributed. Another indication for normality of this group are histogram and QQ-Plot of the modified data in Figure 6. Both of the plots suggest that the modified box counting dimension of mammary cancer tissue is normal distributed.

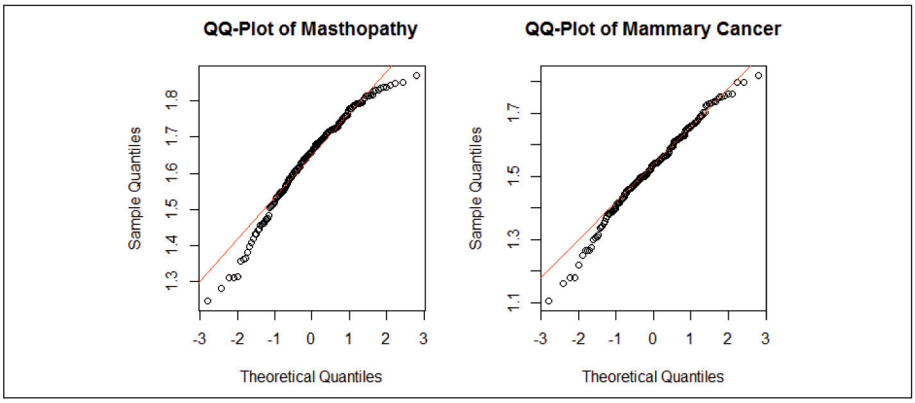


Figure 5. QQ-Plot of the groups masthopathy (left) and mammary cancer (right).

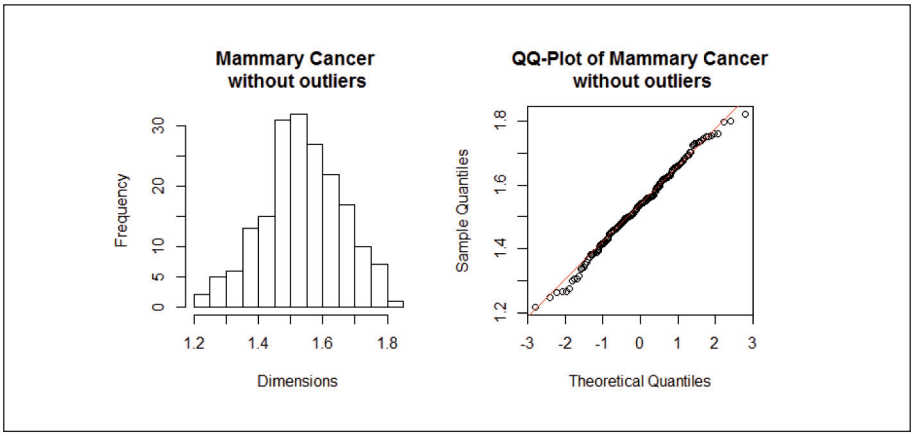


Figure 6. Histogram and QQ-Plot of mammary cancer data without outliers (n = 188).

Robust normality testing procedures have been applied to both groups. Therefore, data has been truncated in the lower boundaries, such that only tissue higher than threshold ε has been taken into account. Shapiro-Wilk tests have been used to compute p-values for the fit of normal distribution. The development of p-values can be found in Figure 7, where especially modified box-counting dimensions of mammary cancer tissue can be seen as normally distributed in contrast to mastophatic tissue box-counting dimensions.

The discrimination between the groups in terms of distributional fit allows the interpretation that more abnormal tissue follows normal distribution. Moreover, it has been shown that Gamma as well as Weibull distributions are proper distributions for fitting mammary as well as masthopathic box-counting dimensions.

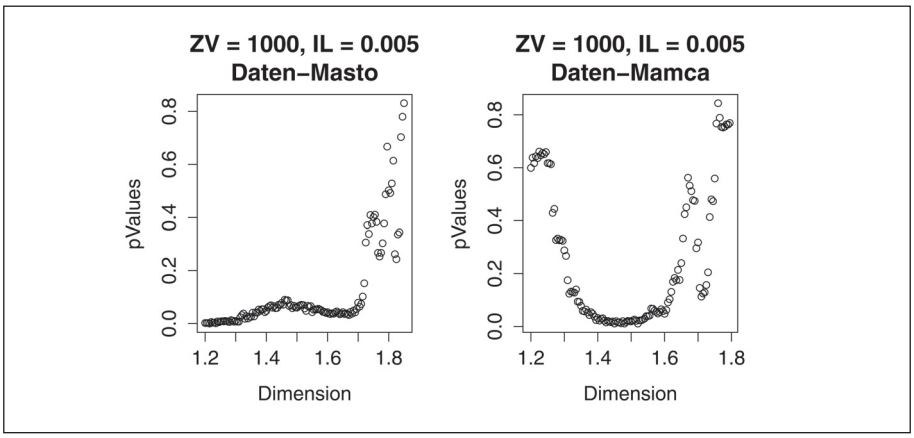


Figure 7. Test for nomal distribution with truncated data for both groups.

REFERENCES

Baish, J.W., and Jain, R.K. (2000). "Fractals and Cancer". *Cancer Res.*, 60: 3683-3688.

George, L.E., and Kamal, H.S. (2007). "Breast Cancer Diagnosis Using Multifractal Dimension Spectra". *2007 IEEE International Conference on Signal Processing and Communications* (ICSPC 2007).

Hermann, P., Piza, S., Ruderstorfer, S., Spreitzer-Groebner, S., and Stehlík, M. (2014). "Fractal Case Study for Mammary Cancer, Biometrie und Medizinische Informatik". *Greifswalder Seminarberichte*, 23: 149-166.

Mrkvicka, T., and Mattfeldt, T. (2011). "Testing *histological images* of *mammary tissues* on Compatibility with the Boolean Model of Random Sets". *Image Anal Stereol* 2011; 30: 11-18.

Paseka, J., Solovyov, S. A., and Stehlík, M. (2015). "Lattice-Valued Bornological Systems". *Fuzzy Sets and Systems*, 259, 15, 68-88

Paseka, J., Solovyov, S. A., and Stehlík, M. (2014). "On the Category of Lattice-Valued Bornological Vector Spaces". *Journal of Mathematical Analysis and Applications,* 419, 138-155.

Stehlík, M. (2015). "On Convergence of Topological Aggregation Functions". *Fuzzy Sets and Systems*.

89. NOTES ON ROBUST TESTING FOR NORMALITY IN INSURANCE SCIENCE

L. Střelec
Mendel University in Brno, Faculty of Business and Economics, Department of Statistics and Operation Analysis, Czech Republic, lubos.strelec@mendelu.cz

ABSTRACT

Insurance science is known as the discipline that applies mathematical and statistical methods. Within those, fitting of claims distribution is a very interesting and important issue. As it is generally known, the claim amounts to the insurance company can be usually described by several continuous non-negative random variables. Moreover, most insurance data is skewed to the right and therefore such probability distributions as exponential, lognormal, Weibull, Pareto, gamma and Burr distributions are usually used. The normal distribution is also important distribution used in insurance and risk management since it usually appears like limiting distribution in many cases. Similarly, many stochastic actuarial models, classical statistical tests and confidence intervals, widely used in insurance, require that the data has been generated by a normal distribution. Thus, the aim of this contribution is to present and discuss some interesting results of robust testing for normality with applications in insurance. Two illustrative examples using claims incurred data and the theft claims data made in a household insurance portfolio are also presented.

1. INTRODUCTION

Insurance science is the discipline where mathematical and statistical methods are extensively used. One of the important issue in insurance is fitting of claims distribution, because it serves an important tool for the purpose of pricing insurance

products. Consequently, it is also possible to consider that the claim amounts to the insurance company are a realization of a probabilistic process which can be fitted by several non-negative right skewed probability distributions, such as exponential, log-normal, Weibull, Pareto, gamma and Burr distributions. However, the normal distribution is also frequently used in insurance. It usually appears like limiting distribution in many cases. For example, the normal approximation should be used for the calculation of aggregate claims distributions (see Dickson, 2005), because it is expected that the distribution of the sum of a number of i.i.d. random variables to tend to a normal distribution by the Central Limit Theorem.

Similarly, many stochastic actuarial models, classical statistical tests and confidence intervals, widely used in insurance and risk assessment, require that the data has been generated by a normal distribution. Testing for normality could be therefore an important topic in these fields. Many common used tests for normality are however sensitive to small deviations from normality. It is therefore often better to use tests from the little-known class of robust tests for normality. In this context, as a robust test for normality it is called the normality test that is less sensitive to small deviations from normality, particularly in the form of a few outliers (Stehlík *et al.*, 2015). Thus, the aim of this contribution is to present and discuss some interesting results of robust testing for normality. The rest of paper is organized as follows. In the second part, the general RT class of tests for normality is introduced. In the third part, we present and discuss some interesting results of power of selected classical and robust tests for normality against distributions usually used in insurance. In the next part, two illustrative examples using claims incurred data and the theft claims data made in a household insurance portfolio are presented. Finally, the last section is Summary.

2. GENERAL RT CLASS TESTS FOR NORMALITY

The general RT class is based on robustification of the classical Jarque-Bera test introduced by Jarque and Bera (1987). The general RT class test statistic is defined by Stehlík *et al.* (2012) for purpose of robust testing for normality against Pareto tails and has the following form:

$$RT = \frac{N}{C_1}\left(\frac{M_{j_1}^{\alpha_1}(r_1, T_{(i_1)}(s_1))}{M_{j_2}^{\alpha_2}(r_2, T_{(i_2)}(s_2))} - K_1\right)^2 + \frac{N}{C_2}\left(\frac{M_{j_3}^{\alpha_3}(r_3, T_{(i_3)}(s_3))}{M_{j_4}^{\alpha_4}(r_4, T_{(i_4)}(s_4))} - K_2\right)^2,$$

where *Mj* are *j*th theoretical central moment estimators of the random vari able defined as $M_j(r, T(F_n, s)) = \frac{1}{n-2r}\sum_{m=r+1}^{n-r}\varphi_j(X_{(m)} - T(F_n, s))$ for $j \in \{0, 1, 2, 3, 4\}$, where φ_j is a tractable and continuous function where $\varphi_0(x) = \sqrt{\pi/2}|x|$ and $\varphi_j(x) = x_j$ for $j \in \{1, 2, 3, 4\}$, $X(m)$ is the order statistic, $T(F_n, s)$ is a location functional applied to the sample $X_1, X_2, \ldots, X_n$, r and s are the trimming constants for moments and location, respectively, K_1 and K_2 are small-sample variants of mean corrections, C_1 and C^2 are constants which can be obtained from the Monte Carlo simulations, and finally, α_1, α_2, α^3 and α_4 are exponents.

As it was mentioned above, for the construction of the general RT class test statistic we use a location functional approach. Note, that the location functional has been introduced by P. E. Bickel and E. L. Lehmann in a series of papers (e.g. Bickel and Lehmann, 1975) and it looks to be playing a crucial role also by robust testing. As it was shown in Stehlík *et al.* (2012), the power of RT class test mimics the effectiveness of location estimator in typical cases. Thus, trade-off between power and robustness is a typical issue here. Based on the location functional approach, in the general RT class we use the following four different location estimators:

- mean: $T_{(0)} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- median: $T_{(1)} = F^{-1}(1/2)$;
- trimmed mean: $T_{(2)}(s) = \frac{1}{n-2s}\sum_{i=s+1}^{n-s} X_{(i)}$, where $X_{(i)}$ is the order statistic of the sample and s is the trimming constant for location;
- pseudo-median: $T_{(3)}$ = median$_{i\leq j}\,(X_i + X_j)/2$, i.e. the median of the set $\{(X_1 + X_1)/2, (X_1 + X_2)/2, (X_1 + X_3)/2, \ldots, (X_1 + Xn)/2, (X_2 + X_2)/2, (X_2 + X_3)/2, \ldots, (X_2 + X_n)/2), \ldots, (X_{n-1} + X_n)/2, (X_n + X_n)/2\}$.

Note, that some theoretical results on consistency and asymptotic distribution χ^2 of the general RT class test statistic can be found in Stehlík *et al.* (2012, 2015).

3. POWER SIMULATION STUDY

To compare the power of several classical and robust tests for normality, we perform the Monte Carlo simulation study using R (R Core Team, 2014). As alternative distributions we consider distributions common used in insurance, i.e. exponential, Pareto, Weibull, log-normal, gamma and Burr distributions. These distributions could be separated in two families – light and heavy-tailed distributions, while the exponential distribution is the border between them. In our study gamma distribution and Weibull distribution with shape parameter $k \geq 1$ represent light-tailed distributions, while the Weibull distribution with shape parameter $0 < k < 1$, and log-normal, Pareto and Burr distributions represent heavy-tailed distributions.

For our simulation study we consider some classical non-robust tests of normality with higher power against the broad scale of alternative distributions – the Shapiro-Wilk test (SW, Shapiro and Wilk, 1965) as the most popular omnibus test of normality for a general use, the Jarque-Bera test (JB, Jarque and Bera, 1987) as the most widely adopted omnibus test of normality in finance and related fields, and the Anderson-Darling test (AD, Anderson and Darling, 1952) and the Lilliefors test (LT, Lilliefors, 1967) as the most famous tests of normality based on the empirical distribution function – accompanied with several new tests for normality based on robust characteristics, in particular, the medcouple test (MC-LR) introduced by Brys *et al*. (2008), the robust Jarque-Bera test (RJB) introduced by Gel and Gastwirth (2008), and the selected robust tests from the RT class, namely MMRT1, MMRT2, TTRT1 and TTRT2 – for more details of these RT class tests see Stehlík *et al*. (2015). The number of replications of the Monte Carlo simulations is 100, 000 and analyzed sample sizes are $n \in \{20, 100\}$.

Table 1 presents the power of mentioned classical and robust tests for normality against alternatives usually used in insurance. From the simulation study we can conclude that SW, AD, MMRT1 and MMRT2 are the most powerful tests. We can also conclude that power of robustified Jarque-Bera tests MMRT1 and MMRT2 for small sample size $n = 20$ is reasonably higher than power of the classical JB test. Power of these tests for large sample sizes is comparable. Therefore, MMRT1 is preferable against this class of alternative distributions when a robust test is desired, especially for small sample sizes.

n = 20	exp (1)	gamma (2,1)	Weibull (2,1)	Weibull (0.5,1)	lnorm (0,1)	Pareto (2,1)	Burr (2,2,1)
AD	0.773	0.460	0.129	0.997	0.903	0.965	0.569
JB	0.624	0.405	0.122	0.959	0.823	0.902	0.549
LT	0.584	0.325	0.104	0.983	0.792	0.901	0.438
RJB	0.558	0.354	0.104	0.943	0.784	0.875	0.509
SW	0.833	0.526	0.152	0.999	0.930	0.978	0.624
MC-LR	0.384	0.167	0.073	0.920	0.482	0.663	0.173
MMRT1	0.722	0.457	0.134	0.990	0.880	0.946	0.578
MMRT2	0.709	0.440	0.129	0.989	0.872	0.943	0.561
TTRT1	0.354	0.241	0.105	0.419	0.361	0.358	0.243
TTRT2	0.534	0.321	0.086	0.944	0.774	0.872	0.476
n = 100	exp (1)	gamma (2,1)	Weibull (2,1)	Weibull (0.5,1)	lnorm (0,1)	Pareto (2,1)	Burr (2,2,1)
AD	1.000	0.998	0.611	1.000	1.000	1.000	0.999
JB	1.000	0.995	0.553	1.000	1.000	1.000	0.999
LT	1.000	0.954	0.392	1.000	1.000	1.000	0.983
RJB	1.000	0.979	0.437	1.000	1.000	1.000	0.996
SW	1.000	1.000	0.787	1.000	1.000	1.000	1.000
MC-LR	0.956	0.631	0.172	1.000	0.989	0.999	0.645
MMRT1	1.000	0.989	0.536	1.000	1.000	1.000	0.997
MMRT2	1.000	0.989	0.548	1.000	1.000	1.000	0.997
TTRT1	0.996	0.877	0.366	1.000	1.000	1.000	0.950
TTRT2	0.999	0.950	0.322	1.000	1.000	1.000	0.988

Table 1. Power of selected normality tests, n = 20 and n = 100.

4. ILLUSTRATIVE EXAMPLES

For illustration purposes, we consider two insurance data sets. The first data set is representing by claims incurred (net of reinsurance) data in the Czech Republic in the period 2002–2012 (data set: CZSO, 2014). The second data set is representing by the theft claims data made in a household insurance portfolio, which are taken from Table 2.1 in Boland (2007). For this case, Boland (2007, p. 48) states that the results of fitting of the original theft claims data give some support for using a log-normal distribution to model this data set. As it is generally known, if the random variable X is log-normally distributed, then $Y = \log(X)$ has a normal distribution. Thus, we also apply mentioned normality tests for the logarithm data to verify this claim.

For mentioned data sets, in Table 2 test statistics and p-values of the Shapiro-Wilk test (SW), the Jarque-Bera test (JB), the Anderson-Darling test (AD), the Lilliefors test (LT), the medcouple test (MC-LR), the robust Jarque-Bera test (RJB), and four robust tests from the RT class are presented.

	Claims incurred		Theft		log(Theft)	
	statistic	p-value	statistic	p-value	statistic	p-value
AD	0.109	0.998	18.6	0.000	0.707	0.064
JB	0.487	0.700	5538.7	0.000	9.776	0.017
LT	0.096	0.995	0.3	0.000	0.087	0.028
RJB	0.335	0.758	394035.3	0.000	13.449	0.012
SW	0.982	0.971	0.5	0.000	0.980	0.065
MC-LR	0.125	0.981	60.9	0.000	4.808	0.162
MMRT1	0.485	0.706	5259.1	0.000	8.903	0.016
MMRT2	0.486	0.720	3677.9	0.000	8.133	0.018
TTRT1	0.227	0.904	42.1	0.000	1.079	0.601
TTRT2	0.162	0.908	334162.4	0.000	5.613	0.063

Table 2. Test statistics and p-values for claims data.

The results presented in Table 2 show, that normality of claims incurred data is not rejected by any of the given tests. These results are in accordance to the assumption that it is usually possible to model claims as normal variables in a short-terms and log-normal in a long-terms (Stehlík and Střelec, 2008). If we consider the theft claims data, so the hypothesis of a normal distribution of log(Theft), i.e. the hypothesis of a log-normal distribution of origin theft claims data, supports the AD, SW, MC-LR, TTRT1 and TTRT2 tests, at 5% significant level. Note that only MC-LR, TTRT1 and TTRT2 tests have good robustness properties (see Stehlík *et al.*, 2015). In the other cases, the null hypothesis is rejected. As it was also shown in Stehlík *et al.* (2015), TTRT1 test is preferable when a robust test is desired, if we consider large sample sizes and heavy-tailed alternatives. Thus, we can conclude that the theft claims data made in a household insurance portfolio could be fit by a log-normal distribution to model this data set.

5. CONCLUSION

In this contribution some interesting results of robust testing for normality were presented and discussed. Firstly, the general RT class test statistic was introduced and the results of power of selected classical and robust tests against alternative distributions usually used in insurance were presented. We can conclude that power of selected RT tests is reasonably higher than power of the classical JB test, especially for small sample sizes. In the last section of this contribution two illustrative examples were presented. In accordance to Stehlík *et al*. (2015) we can conclude that if we primarily focus on tests which have good robustness properties for large sample sizes, the TTRT1 test is preferable.

ACKNOWLEDGMENTS

Special Thanks to Dr. Stehlík for cooperation on research of robust testing for normality.

REFERENCES

Anderson, T.W., and Darling, D.A. (1952). "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes". *The Annals of Mathematical Statistics*, 23, 193-212.

Bickel, P.J., and Lehmann, E.L. (1975). "Descriptive Statistics for Nonparametric Models II. Location". The Annals of Statistics, 3, 1045-1069.

Boland, P.J. (2007). *Statistical and Probabilistic Methods in Actuarial Science*. London: Chapman and Hall/CRC.

Brys, G., Hubert, M., and Struyf, A. (2008). "Goodness-of-Fit Tests Based on a Robust Measure of Skewness". *Computational Statistics*, 23, 429-442.

CZSO. (2014). *Economic Results of Insurance Companies, Pension Companies and Funds*, http://www.czso.cz/csu/2014edicniplan.nsf/engpubl/042007-14-eng_r_2014.

Dickson, D.C.M. (2005). *Insurance Risk and Ruin*. Cambridge: University Press.

Gel, Y.R., and Gastwirth, J.L. (2008). "A Robust Modification of the Jarque-Bera Test of Normality". *Economics Letters*, 99, 30-32.

Jarque, C.M., and Bera, A.K. (1987). "A Test for Normality of Observations and Regression Residuals". *International Statistical Review*, 55, 163-172.

Lilliefors, H.W. (1967). "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". *Journal of the American Statistical Association*, 62, 399-402.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Shapiro, S.S., and Wilk, M.B. (1965). "An Analysis of Variance Test for Normality". *Biometrika*, 52, 591-611.

Stehlík, M., Fabián, Z., and Střelec, L. (2012). "Small Sample Robust Testing for Normality against Pareto Tails". *Communications in Statistics – Simulation and Computation*, 41, 1167-1194.

Stehlík, M., and Střelec, L. (2008). "On Normality Assumptions for Claims in Insurance". *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 57(3), 141-146.

Stehlík, M., Střelec, L., and Thulin, M. (2015). "On Robust Testing for Normality in Chemometrics". *Chemometrics and Intelligent Laboratory Systems*, 130, 98-108.

90. AN INTRODUCTION TO THE SAFETY ASSESSMENT IN KOREA OF CHEMICALS IN HOUSEHOLD PRODUCTS

J.H. Suh, M.S. Hwang, H.S. Shin, Y.J. Seo and I.G. Hwang
Food Safety Risk Assessment Division, National Institute of Food and Drug Safety Evaluation, Ministry of Food and Drug Safety, Korea,
hwang1963@korea.kr

H.J. Yoo
Department of Consumer Studies,Chungbuk National University, Korea
yoohj@chungbuk.ac.kr

J.H. Lee
Institute of Environmental Safety and Protection,
NeoEnBiz Co., Korea
jhlee@neoenbiz.com

ABSTRACT

Household products contain various chemicals and their potential risk to the human health is ever present. Safety concerns of household products have been raised with humidifier sterilizer incident in Korea 2011. Based on this issue, Korea National Institute of Food and Drug Safety Evaluation (NIFDS) selected 28 chemicals widely used in detergent and wet tissues and performed safety assessment of these chemicals. In this study, the exposure algorithms were designed specifically for Korean population and coefficient factors were determined from a consumer survey conducted on 3,000 Korean population. The margin of exposure for each chemical was calculated by comparing the NOAEL to the estimated human exposure. We have concluded that risk of chemical exposed to human body from the use of wet tissue is negligible. However, the use of detergent was associated with high exposure level of ethoxylated alcohols, coconut deiethanolamide and limonene. We recommend

the public to wear protection equipment when using detergent that can come in direct contact with human body.

This study will be applicable to establish the basis of risk assessment and to strengthen safety management system in Korea.

1. INTRODUCTION

The use of household products such as detergent and wet tissue has dramatically increased in recent years due to convenience and ease of use. These products contain numerous chemical substances as surfactant for removing stain or dirt, preservative to prevent microbial contamination, aroma components and etc. Nevertheless, safety of each chemical present in frequently used household products not yet clearly defined. In addition, safety is not guaranteed by labeling a product as dangerous and not using it simply because it contains an ingredient with dangers.

In 2011, dozens of pregnant and postpartum women have simultaneously developed interstitial lung diesease from unknown cause and died. This incident and subsequent reports described the high prevalence of this disease in pregnant and postpartum women and their children, and its 100% mortality rate. These serial reports alerted the Korea Centers for Disease Control and Prevention (KCDC), which initiated investigations to identify the cause of the disease and come up with preventive measures. On August 31, KCDC announced to the media that their preliminary epidemiological analyses of acute lung injury with unkown causes in adult cases revealed a strong association with the use of humidifier disinfectants (Yang *et al*., 2013). KCDC also reported PHMG (Poly hexa methylene guanidine) and PGH (Oligo (2-(2-ethoxy) ethoxy ethyl guanidinium chloride) as a inhalation toxicity agent of the humidifier disinfectants. This issue had great social repercussions and also alerted government bodies confirming safety concern of household products. Also public safety concerns of chemicals in household products used in everyday lives have increased.

The Korean government has decided to conduct a safety assessment of household products which may pose potential harm to korean consumers.

The aim of this study was to conduct safety assessment of 5 groups of household products including synthetic detergent, fabric softener, bleaching agent, detergent and wet tissues frequently used by Korean consumers**.** To determine the risk associated with the use of such products, it is necessary to grasp level of chemical exposure from the use of a product and whether such chemical poses potential harm to human health.

Korea NIFDS surveyed 3,000 Korean consumers to identify uses patterns of household products and conducted exposure assessment through established algorithms of each group of products. This study provides scientific background in safety management system for synthetic detergent, fabric softener, bleaching agent, detergent and wet tissues in Korea.

2. OUR APPROACH

Chemical substances selection criteria

We have collected the lists of chemicals used in 413 detergent and wet tissues from Korean Agency for Technology and Standards (KATS)[1] , industries and labeling of the products. Through the exclusion of repetition and unconfirmed substances, 400 chemical substances were summarized.

182 GRAS (Generally Recognized as Safe) chemicals with unsufficient toxiological information not being assessed in this study. Based on hazard potential, urgency social concerns, and frequency of use in products, total of 28 chemical substances were selected for risk assessment in this study (Table 1).

[1] KATS: KATS has strengthened policy activities in standards and product safety for industrial product in Korea. This agency have managed registered information of ingridient substaces used in wet tissues, detergent etc..

	Purpose	Chemicals	CAS No.
1		Methylisothiazolinone	2682-20-4
2		Chloromethylisothiazolinone	26172-55-4
3		Ethanol	64-17-5
4		Poly(hexamethylenebiguanide)hydrochloride	32289-58-0
5		Iodopropynyl Butylcarbamate	55406-53-6
6	Preservative	1,2-Benzisothiazolin-3-one	2634-33-5
7		Dimethyloldimethyl Hydantoin	6440-58-0
8		(Tallow)Trimethylammonium Chloride	8030-78-2
9		Coconut Diethanolamide	68603-42-9
10		Propyl Paraben	94-13-3
11		2-Bromo-2-nitro-1,3-propanediol	52-51-7
12		Ethylene Oxide	75-21-8
13		Thiourea	62-56-6
14		Alcohols, C12-14, ethoxylated	68439-50-9
15		Alcohols, C9-11, ethoxylated	68439-46-3
16		Alcohols, C16-18, ethoxylated	68439-49-6
17		Alcohols, C12-13, ethoxylated	66455-14-9
18	Surfactant	Decyl alcohol, ethoxylated	5131-66-8
19		1-Butoxy-2-Propanol	26183-52-8
20		Fatty alcohol polyoxyethylene ether	9002-92-0
21		Sodium Lauryl Ether Sulphate	9004-82-4
22		Diethanolamine	111-42-2
23	Aroma component	Limonene	5989-27-5
24	Abrasive	Quartz	14808-60-7
25	Foam	N,N-Dimethyldodecylamine-N-oxide	1643-20-5
26	stabilizer	Sodium Metasilicate Nonahydrate	13517-24-3
27	Softener	Dioctyl Phthalate	117-84-0
28	Viscosity agent	Ammonium Thioglycolate	5421-46-5

Table 1. 28 chemical substances.

Exposure assessment

The exposure of chemical substance to consumers is determined mainly by the frequency and exposure time and its concentration in the household products. It is necessary to gain an understanding of how a product is used by consumer to establish

coefficient factor. Thus, in this study a survey of 3,000 Korean consumers who used more than 1 detergent or wet tissues within a month was conducted. Consumers feed information about demographics (age, sex, etc.), type and amount of household product used, frequency and etc. through online program during 13 days (Yoo *et al.*, 2014 a, b). We have classified household products found in Korean market into 26 groups and then set exposure scenarios. Based on The Soap and Detergent Association (2005) and Nethalands National Institute for Public Health and the Environment (2006) studies, the algorithm for each exposure scenario was revised and complemented to fit Korean situation. And also each substance physicochemical and toxicological characteristics and surface area were obtained using references (WHO report 1993-2003, OECD report 2004~2009 and etc.). Finally, exposure algorithms and coefficient factors adjusted for Korean population were developed (Table 2).

Dermal Exposure	
Direct	$D = A_{skin} \times W_f \times n \times Abs \div BW$
Continuous	$D = A_{time} \times W_f \times t_r \times n \times Abs \div BW$
Transition from fabric	$O = A \times W_f \times Abs \times S_{der} \times FD \times F1 \div W \times F2 \times F3 \times F4 \times n \div BW$
Inhalation Exposure	
Direct	$C = {}_{Adisperse} \times W_f + V \times e^{-qt} \times n \times t_r \div 24$
Volatiles	$C = [-(A \times W_f \div t_r) \div (q \times V) \times 1 - e^{(-qtr)} \div q \times e^{(-qt)} + (A \times W_f \div tr) \div (q \times V) \times 1 - e^{(-qtr)} \times e^{(qtr)} \div q] \times F_{air} \times n \div 24$
Aerosol	$C = f_{airborne \leq 20\mu m} \times (M \times W_f \times t_r) \div (V \times q) \times (1 - e^{(-qtr)}) \times n \div 24$
Oral Exposure	
General type	$D = A_{intake} \times W_f \times n \div BW$

A: Amount of using product per once (mg/frequency) A_{time}: Residue of skin per time (mg/min/frequency) A_{intake}: Amount of intake per once (mg/frequency) n: Frequency per day (frequency/day) t: Exposure time (min) FD: Fabric density (mg/cm^2) F2: Transition from fabric proportion F4: Residue proportion of detergent after final dehydration V: Room volume (m^3) M: Spraying amount per time (mg/min/frequency) $f_{airborne \leq 20\mu m}$: Airborne fraction($\leq$20μm)	A_{skin}: Residue of skin per once (mg/frequency) $A_{disperse}$: Amount of dispersed product (mg/frequency) W_f: Composition of product (%) Abs: Absorption rate (%) t_r: Using time (min) S_{der}: Contact area (cm^2) F1: Residue proportion of detergent at last washing F3: Proportion of residue in skin W: Laundry weight (mg) q: Ventilation rate (1/min) F_{air}: Air release factor

Table 2. Algorithms for each exposure route.

Margin of safety

The amount of exposure was calculated using algorithms for each scenario. The NOAEL is compared to the various estimates of human exposure to calculate the margin of exposure (MOE). The magnitude of the MOE gives an indication of the level of concern, but is not a precise quantification of risk (Benford *et al*., 2010).

3. CONCLUSION

The safety assessments for 28 chemicals widely used in detergent and wet tissues were conducted in this study. We have concluded that risk of chemical exposed to human body from the use of wet tissue is negligible. However, the use of detergent was associated with high exposure level of ethoxylated alcohols, coconut deiethanolamide and limonene. Korea NIFDS recommends establishing safety management system to indicate safety instruction on product label (for example, wear protection equipment when using detergent). In Korea, safety regulations of detergent, synthetic detergent, fabric softener and bleaching agent are legislated by Ministry of Environment and those of wet tissues are controlled by KATS and MFDS.

ACKNOWLEDGEMENTS

This research was supported by a grant 13161MFDS784 and 13162MFDS785 from Ministry of Food and Drug Safety in 2014.

REFERENCES

Choi, YY., Lim, HK., Lim, SY., and Paek, DM. (2012). "Health Damages and Lessons of the Use of Humidifier Disinfectants in Korea". *J Environ Health Sci* 2012, 38 (2), 166-174.

Benford, D., Bolger, P.M., Carthew, P., Coulet, M., DiNovi, M., Leblanc, J.D., and Wildemann, T. (2010). "Application of the Margin of Exposure (MOE) Approach to

Substances in Food that Are Genotoxic and Carcinogenic". *Food and Chemical Toxicology*, 48, S2-S24.

Korea Centers for Disease Control and Prevention (2011). "Interim Report of Epidemiological Investigation on Lung Injury with Unknown Cause in Korea". *Public Health Wkly. Rep.*, 4, 817-818.

Netherlands National Institute for Public Health and the Environment (RIVM) (2006). *Cleaning Products Fact Sheet to Assess the Risks for the Consumer*. RIVM report 320104003/2006.

The Soap and Detergent Association (SDA). (2005). *Exposure and Risk Screening Methods for Consumer Product Ingredients*. SDA.

Yang, HJ., Kim, HJ., Yu, J., Lee, E., Jung, YH., Kim, HY., Seo, JH., Kwon, GY., Park, JH., and Gwack, J. (2013). "Inhalation Toxicity of Humidifier Disinfectants as a Risk Factor of children's Interstitial Lung Disease in Korea: A Case-Control Study". *PLoS One*, 8 (6), e64430.

Yoo, H.W., and Hwang, H.S. (2014a). "Use Pattern and Safety Pursuit Behavior of Consumers who Using Wet Wipes for Infants". *Korean Review of Crisis & Emergency Management*, 10 (9), 49-73.

Yoo, H.W., Hwang, H.S., and Song E.G. (2014b). "A Study on the Improvement for Safe Consumption of Wet Tissue – Focusing on Consumer's Subjective Safety Assessment and Labelling Condition". *Korean Review of Crisis & Emergency Management*, 10 (8), 81-97.

91. CHILDHOOD OBESITY AS CARDIOVASCULAR DISEASE RISK PREDICTOR

L. Tarro, E. Llauradó, M. Aceves-Martins
Universitat Rovira i Virgili, Medicine and Health Science Faculty, NFOC group, Health Education and Promotion, Spain, lucia.tarro@urv.cat, elisabet.llaurado@urv.cat, magaly.aceves@estudiants.urv.cat

D. Moriña
Centre for Research in Environmental Epidemiology (CREAL), dmorina@creal.cat

M. Giralt
Universitat Rovira i Virgili, Medicine and Health Science Faculty, NFOC group, Pharmacobiology group, Spain, montse.giralt@urv.cat

R.Solà
Universitat Rovira i Virgili, Medicine and Health Science Faculty, NFOC group, Research Unit of Lipids and Arteriosclerosis, Servei de Medicina Interna, Hospital Universitari Sant Joan de Reus, IISPV, CIBER Diabetes and Associated Metabolic Disorders (CIBERDEM), Spain, rosa.sola@urv.cat

ABSTRACT

Obesity is a well-known risk factor for several cardiovascular diseases. The increase of childhood obesity in recent years is likely to imply an increase of the prevalence of these overweight and obesity related cardiovascular diseases. Several interventions have been carried out in the last years aiming at revert the obesity prevalence-

increasing trends and to popularize healthy lifestyle including food habits among children. In the present work, Educació en alimentació (EdAl) school-based program is presented: the original intervention at Reus and surroundings towns as control, and the reproduction of this program at "Terres de l'Ebre". These interventions showed the promising results at the end-of-intervention and 2-y follow-up after cessation of intervention. EdAl program achieved a reduction of obesity prevalence and an improvement of lifestyles such as ≥4 after-school physical activity h/week and dietary habits (fruit and vegetable consumption). In conclusion, healthy lifestyles improved due to EdAl program intervention, are identified risk factors of childhood obesity and cardiovascular disease. These results, according to the current evidence, could represent a 4% to 19% risk reduction of cardiovascular disease.

1. INTRODUCTION

Obesity (OB) is a well-known risk factor for cardiovascular diseases (Bastien, Poirier, Lemieux and Després, 2014). There is a strong association between childhood obesity and early-onset comorbidities that have an adverse effect on vasculature, such as hypertension, dyslipidaemia and insulin resistance, which can result in a premature and accelerated artheroschlerosis (Raghuveer, 2010). The increase of childhood obesity rates in recent years is likely to imply an increase of the prevalence of these overweight and OB related cardiovascular risk factors (Boden and Salehi, 2013). A Danish cohort study of 276,835 children found that higher childhood BMI values elevated the risk of having a coronary heart disease event in adulthood. This association is stronger in boys than in girls and increase with the age of the child in both sexes (Baker, Olsen and Sørensen, 2007). The Harvard Growth Study showed that being overweight in adolescence resulted in more than a 2-fold higher relative risk of coronary artery disease mortality, independent of adult weight (Must, Jacques, Dallal, Bajema and Dietz, 1992). Taking this into account, several interventions have been carried out in the last years on children with the objective of revert the OB prevalence-increasing trend and to popularize healthy lifestyle and food habits among them (Verrotti, Penta, Zenzeri, Agostinelli and De Feo, 2014).

According to the European Heart Network (2011), changing some life and health habits could benefit and protect health from Cardiovascular Disease (CVD), were

eating more pieces of fruit or vegetables per day, or practicing a specific time of physical activity could reduce this risk (Table 1).

Fruit and Vegetable Consumption	
Portions (106 gr) 7 day	Risk of CVD reduction
1 portion /day	4%
3-5 portions/day	7%
> 5 portions/day	17%
Physical Activity	
Walking	Risk of CVD reduction
30 minutes/day/5 days a week	19%

The necessity to tackle childhood OB drove an initiative to create a healthy promotion of lifestyles program in primary schools, developed by undergraduate university students formed as Health Promoter Agents (Giralt, Albaladejo, Tarro, Moriña, Arija and Solà, 2011).

This successful intervention was carried out in two different locations in southern Catalonia, Reus and surroundings towns (Tarro, Llauradó, Albaladejo, Moriña, Arija and Solà, 2014), and "Terres de l'Ebre" (Llauradó, Tarro, Moriña, Queral, Giralt and Solà, 2014) proving the reproducibility of the program and the promising results at the end-of-intervention and 2-y followup after cessation of intervention (Tarro, Llauradó, Moriña, Solà and Giralt, 2014). EdAl program improves healthy lifestyles that reduce childhood OB, and these improvements are maintained after the intervention cessation in primary-school children. The purpose of this work was to examine the association between risk factors of obesity and their improvement over the EdAl programs progression.

2. PROTECTIVE AND RISK FACTORS OF CHILDHOOD OB IN EDAL PROGRAMMES

The risk factors of childhood OB were calculated by odds ratio (ORs) for dietary and physical activity patterns of EdAl participants. We refer to Table 1 for a summary of protective and risk factors of childhood obesity in EdAl programs. We refer to Table 2 for a summary of protective and risk factors lifestyles improved over EdAl programmes.

As we can observe, some of the identified protective and risk factors of childhood OB were improved by EdAl programmes. The physical activity is the most improved lifestyle in both programmes and follow-up study. However, the program that presents the best improving in identified protective and risk factors against childhood OB, is EdAl-2 (Llauradó, Tarro, Moriña, Queral, Giralt and Solà, 2014), the reproduction of EdAl program in "Les Terres de l'Ebre". Consequently, if there are improvements of protective and risk factors of childhood OB we hope that in future follow-up assessments the OB and cardiovascular risk factors will be reduced.

	Protective Factor of OB	Risk Factor of OB
EDAL-1[1]		
Fish consumption	OR=0.390*	
Fast-food consumption		OR=2.270*
EDAL-2[2]		
Dairy product at breakfast	OR=0.336*	
≥ 4 after-school PA h/week	OR=0.600*	
< 4 after-school PA h/week		OR=1.811*
EDAL-1 2-y follow-up[3]		
≥ 4 after-school PA h/week	OR=0.240*	
Daily fruit consumption	OR=0.447*	
≤ 2 after-school PA h/week		OR=2.203*

Table 1. Protective and Risk Factors of childhood OB.
[1] Tarro, Llauradó, Albaladejo, Moriña, Arija and Solà, 2014
[2] Llauradó, Tarro, Moriña, Queral, Giralt and Solà, 2014)
[3] Tarro, Llauradó, Moriña, Solà and Giralt, 2014
*$p<0.05$

	EDAL-1[1]	EDAL-2[2]	EDAL-1 2-follow-up[3]
Fish consumption	–	–	–
Fast-food consumption	–	Reduction of -8.3% in intervention group	–
Dairy product at breakfast	–	Reduction of -4.3% in intervention group	–
≥ 4 afterschool PA h/ week	Increase of 8.3% in intervention group	Reduction of 11,8% in intervention group	Increase of 24.9% in intervention group
< 4 afterschool PA h/ week	–	Decrease of 6.6% <2 afterschool PA h/week in intervention group	–
Daily fruit consumption	–	Increase of 11,1% 2nd fruit a day in intervention group and decrease of -8.6% 2nd fruit a day in control group	–
≤ 2 afterschool PA h/ week	–	Decrease of 6.6% <2 afterschool PA h/week in intervention group	–

Table 2. Protective and Risk Factors lifestyles improved over EdAl programmes.
[1] Tarro, Llauradó, Albaladejo, Moriña, Arija and Solà, 2014
[2] Llauradó, Tarro, Moriña, Queral, Giralt and Solà, 2014)
[3] Tarro, Llauradó, Moriña, Solà and Giralt, 2014

3. CONCLUSION

EdAl program achieved a reduction of obesity prevalence and an improvement of lifestyles such as ≥4 after-school physical activity h/week and dietary habits (fruit and vegetable consumption). The healthy lifestyles improved due to EdAl program intervention, are identified as risk factors of childhood obesity and cardiovascular disease. In conclusion, these results, according to the current evidence, could represent a 4% to 19% risk reduction of cardiovascular disease.

ACKNOWLEDGEMENTS

This research project has been supported by Fundació Privada Reddis, Ajuntament de Reus, Spain; Conselleria de Salut de Generalitat de Catalunya; Mercat Central de Reus, Protected Designation of Origin Siurana, Spain [DOP Siurana], La Morella Nuts, S.A. Spain; Nutrition and Health Technology Centre-TECNIO CT09-1-0019; Reus (Spain), Diputació de Tarragona, Spain; Ajuntament d'Amposta, Spain. We express our appreciation to the university medical and health science students of the Facultat de Medicina i Ciències de la Salut, Universitat Rovira i Virgili (Reus, Spain) as well as the staff and parents of the pupils of the primary schools of Reus, Cambrils, Salou, Vila-seca Amposta, Sant Jaume d'Enveja, Els Muntells, l'Ametlla de Mar, El Perelló, l'Ampolla, Deltebre, l'Aldea, Lligalló del Gànguil and Camarles for their enthusiastic support in this study. MA-M is a Mexican recipient of a fellowship award from the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología, CONACyT).

REFERENCES

Baker, J.L., Olsen, L.W., Sørensen, T.I. (2007). "Childhood Body-Mass Index and the Risk of Coronary Heart Disease in Adulthood". *N Engl J Med*., Dec 6; 357 (23): 2329-2337.

Bastien, M., Poirier, P., Lemieux, I., and Després, J.P. (2014). "Overview of Epidemiology and Contribution of Obesity to Cardiovascular Disease". *Progress in* Cardiovascular Disease, 56 (4), 369-381.

Boden, G., and Salehi, S. (2013). "Why Does Obesity Increase the Risk for Cardiovascular Disease?". *Current Pharmaceutical Design,* 19 (32), 5678-5683.

European Heart Network (2011). *Diet, Physical Activity and Cardiovascular Disease Prevention in Europe*. EHN Publications.

Giralt, M., Albaladejo, R., Tarro, L., Moriña, D., Arija, V., and Solà, R. (2011). "A Primary-School-Based Study to Reduce Prevalence of Childhood Obesity in

Catalunya (Spain) – EdAl (Educació en Alimentació) Study: Protocol for a Randomized Controlled Trial". *Trials,* 12: 54.

Llauradó, E., Tarro, L., Moriña, D., Queral, R., Giralt, M., and Solà, R. (2014). "EdAl-2 (Educació en Alimentació) Programme: Reproducibility of a Cluster Randomised, Interventional, Primary-School-Based Study to Induce Healthier Lifestyle Activities in Children". *BMJ Open,* 4 (11), e005496.

Must, A., Jacques, P.F., Dallal, G.E., Bajema, C.J., Dietz, W.H. (1992). "Long-Term Morbidity and Mortality of Overweight Adolescents. A Follow-Up of the Harvard Growth Study of 1922 to 1935". *N Engl J Med.*, Nov 5; 32 7 (19): 1350-1355.

Raghuveer G. (2010). "Lifetime Cardiovascular Risk of Childhood Obesity". *Am J Clin Nutr.*, May; 91 (5): 1514S-1519S.

Tarro, L., Llauradó, E., Albaladejo, R., Moriña, D., Arija, V., and Solà, R. (2014). "A Primary-School-Based Study to Reduce the Prevalence of Childhood Obesity – The EdAl (Educació en Alimentació) Study: A Randomized Controlled Trial". *Trials,* 15: 18.

Tarro, L., Llauradó, E., Moriña, D., Solà, R., and Giralt, M. (2014). "Follow-Up of a Healthy Lifestyle Education Program (the Educació en Alimentació Study): 2 Years after Cessation of Intervention". *Journal of Adolescent Health*, 55 (6): 782-789.

Verrotti, A., Penta, L., Zenzeri, L., Agostinelli, S., and De Feo, P. (2014). "Childhood Obesity: Prevention and Strategies of Intervention. A Systematic Review of School-Based Interventions in Primary-Schools". *Journal of Endocrinological Investigation,* 9 (Epub ahead of print).

92. VALIDATION OF RISK SCORING MODELS FOR TRADE CREDIT IN SMALL AND MEDIUM ENTERPRISES

M. Terradez and A. Debon
Universitat Politècnica de València1, Dept. of Applied Statistics, Operations Research and Quality, Spain, {mterrade, andeau@eio.upv.es}@eio.upv.es

Á.A. Juan
IN3-Open University of Catalonia, Computer Science Dept., Spain, ajuanp@uoc.edu

ABSTRACT

In a previous work the authors analyzed the relationship between the trade credit (TC) management and the level of risk in SMEs from a statistical perspective. This paper is a new step in this direction, as it focuses on the validation of the model from a double point of view: a necessary statistical validation and a complementary validation from the business perspective. Thus, in this study we first show that TC is a good proxy of risk. In order to do that, we have developed a model that explains the binary dependent variable which achieves a high goodness-of-fit. In addition, we have obtained empirical evidence that TC can be an interesting and useful contribution to classical scoring models when dealing with SMEs. In particular, it has been found that the inclusion of variables related to TC can improve the performance of such models.

1. INTRODUCTION

Trade credit (TC) involves supplying goods and services on a deferred payment basis. Its management is a matter of great importance for business, especially for small and medium-sized enterprises (SMEs), since it represents a significant portion of

their assets. In a previous work (Terradez *et al.*, 2015), the authors analyzed the relationship between the TC management and the level of risk in SMEs from a statistical perspective. They used a simplifying approach concerning a binary dependent variable (high risk vs. low risk), and generated a decision tree model able to efficiently order a list of companies according to their risk (what is commonly known as a score). This paper complements our previous work by focusing on the validation of the model from a double point of view. On the one hand, it uses some of the more frequently accepted validation techniques (split of the sample and cross-validation) in order to validate the model. On the other hand, it performs a validation from the business perspective.

The remainder of this paper is structured as follows: Section 2 offers a literature review on related work. Section 3 provides an overview of our validation approach. Section 4 explains the methodology used for the statistical validation. Section 5 focuses on the ordinal model as a business validation. Finally, Section 6 provides some concluding remarks.

2. LITERATURE REVIEW

Seifert *et al.* (2013) provide a comprehensive literature review on economic and financial theories regarding the relevance of TC analysis. Garcia-Teruel and Martinez-Solano (2007) analyze a large sample of Spanish SMEs by using panel data models. They conclude that: *(i)* managers can create value by reducing the average collection period from customers, and *(ii)* reducing the cash cycle increases the profitability of the company. Deloof (2003) derived a similar conclusion from a sample of Belgian companies. Furthermore, Martinez-Sola *et al.* (2012) analyze the effect of TC policies on the stock market value of the company, concluding that there is a concave relationship: positive for low levels of investment in TC, and negative for high levels.

Recent research has focused on the links between the management of TC (and late payment) and the liquidation or bankruptcy of enterprises, or even the refinancing (or restructuring) of debt (Wilner, 2000). Commercial lending involves credit risk due to exposure to default that can have negative effects on probability and liquidity (Cheng and Pike, 2003). However, works that have linked the use of trade credit with

liquidation and insolvency processes have been mostly theoretical. Late payment problems have been described from surveys in the UK (Peel *et al*., 2000) market, with descriptive studies focused on the factors that employers perceive as impediments to improving the performance of their business. Management of TC offered (as a supplier) is critical to the survival and success of business, that SMEs are not proactive in their management of credit and that there is a clear gap in the application of risk models (83% of them do not classify their customers according to risk categories). Recent studies have shown that credit-constrained firms facing liquidity problems from their customers are more likely not to pay their suppliers, and, consequently, a chain of payment difficulties (Boissay and Gropp, 2013). However, because of the difficulty of obtaining data, the line of research that studies the relation between the management of TC and risk is not sufficiently developed.

3. METHODOLOGY OVERVIEW

A common measure of predictive accuracy of trees is the risk. For categorical dependent variables, the risk estimate is the proportion of cases incorrectly classified after corrected with respect to prior probabilities and misclassification costs, while for scale dependent variables, the risk estimate to the variance within the node. Therefore, the lower the value of risk, the greater the accuracy in the estimation. In this work we will use a measure of goodness of fit that has a growing importance when talking about binary models: the ROC curve (Fawcett, 2006). The assessment of the predictive power of the model is defined by the area under the ROC curve (AUC), calculated for all possible cutoff points. Since AUC is an area of the unit square, its value is always between 0 and 1, and the random model has an area of 0.5. The larger the AUC, the better the classifier. Therefore, the AUC will be our main criteria for assessing the goodness of fit of the models. Note that the AUC refers to the correct ranking of companies according to their risk, while the percentage of individuals correctly classified depends obviously on the chosen cutoff point (default is set to 0.5), and therefore it can be adjusted whenever the company wants to follow a risky or a conservative management policy.

The most basic and traditional method for validating a model reserves a percentage of the database as a test set (or validation set). The remaining data are the training

set, which is used to build the model. The division of the data in these two groups should be randomly selected so that the estimation is correct. The training error is the average error that results from applying the model to the observations used in its training, while the test error can be calculated by using the model to predict the response on a new observation that was not used in the training set. Given a data set, the use of a particular model is warranted if it results in a low test error, but the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter. The validation process is thus required to prevent that the accuracy of the model is overestimated by providing much better results for the training set than for the test set (a problem known as over-fitting).

An alternative method of validation is known as cross-validation (CV), which is typically implemented by the method of the *k* folds (*k*-fold cross validation) which randomly divided the data into *k* groups (often *k* = 10) of similar size. A group is reserved as a test set, the model is built with the remaining *k*-1 groups, and then used to predict the outcome of data from the reserved group. This process is repeated *k* times, each time leaving a different group for testing. Finally, we build a model with all the data and obtain their error and/or accuracy rates by averaging the *k* available ratios. CV is a very general approach that can be applied to almost any statistical model. Performing 10-fold CV requires fitting the learning procedure only ten times, which may be much more computationally feasible than other approaches.

4. STATISTICAL VALIDATION

We decided to divide the sample into two subsamples: one for the training of the model and the other one for its validation. Thus, 70% of the sampling observations (6,449 records) were assigned to model training, while the remaining 30% (2,759 records) were kept for validation purposes. Cross-validation divides the sample into a number of disjoint subsamples (or folds) of similar size. Then the tree models are generated: the first tree is based in all cases except those in the first fold of the sample, the second tree is based in all cases except those in the second fold of the sample, etc. For each tree, misclassification risk is calculated by applying the tree to the subsample excluded in generating it. This method generates a

single tree model and a final risk estimate which is calculated as the average of the risks of all trees.

Applying the model obtained in the previous section to the validation sample, we obtained the results shown in Figure 1. According the results, AUC = 0.763 and overall success = 85%.

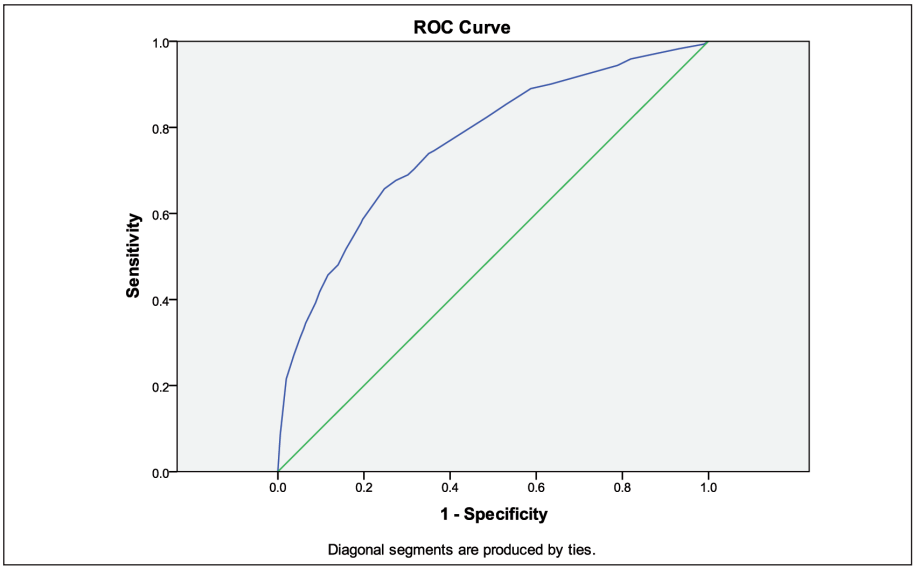


Figure 1. ROC curve in validation.

5. BUSINESS VALIDATION

We will use the variable "MORE_Gr8", which is the original variable "RatingMORE" with a couple of changes: the three highest levels have been grouped, and alphanumeric values are recoded into numerical values with a logical order. We start from a well-known scoring model, the one proposed by Pozuelo *et al.* (2010). We use this model as the main independent variable, and try to assess whether the contribution of the score obtained by our previous model (introduced as an additional independent variable) is statistically significant and helpful in explaining the dependent variable. Therefore, the model used to obtain the score is:

$$Z = -0.692\ REN8 + 12.154\ EF4 - 3.195\ EF8 - 12.152\ SOLVLP2 + 1.304$$

Being REN8 = RO/FP, EF4 = GF/PT, EF8 = FP/PT and SOLVLP2 = RGAI/PT, (RO: Profit, FP: Equity, GF: Financial charges, PT: Total liabilities and RGAI: Profit before taxes). As it is a logit model, the transformation into PD is PD_PLV = 1 / (1 + exp(–Z)).

An ordinal regression model has been carried out, including the score provided by our previous model (variable CHAID) as an independent variable. The parameter associated with the inclusion of this factor in the model is significant, and it improves all measures of goodness of fit, both the Pseudo-R2 and the measures of association of the contingency table.

Goodness-of-fit measure	PD_PLV	PD_PLV+CHAID
-2 Log Likelihood	16811	16506
Pseudo-R2 (Cox-Snell)	0.719	0.735
Pseudo-R2 (Nagelkerke)	0.734	0.750
Pseudo-R2 (Mc Fadden)	0.326	0.341
Correlation (Kendall Tau)	0.744	0.751
Correlation (Gamma)	0.852	0.857
Chi-squared distance	31.543	30.356

Table 1: Comparison of goodness-of-fit measures.

As a result, we consider that adding our previous model as an input improves significantly the scoring model.

6. CONCLUSION

In previous work we developed a model and conclude that TC was a good proxy of risk for SMEs. This work contributes to validate those results from a statistical point of view and also concludes that TC can be an interesting and useful input for classical risk scoring models when dealing with SMEs. In particular, it shows that the incorporation of variables related to TC management can improve the performance of such models. Using an ordinal dependent variable (8 levels) and a known scoring model as independent variable, we have built a new model incorporating the results

of our score as an additional independent variable, deriving that the contribution of our model is statistically significant when trying to explain the dependent variable, and improves the goodness-of-fit measures.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (grants TRA2013-48180-C3-3-P and MTM2013-45381-P) and FEDER. Likewise we want to acknowledge the support received by the Department of Universities, Research & Information Society of the Catalan Government (Grant 2014-CTP-00001).

REFERENCES

Boissay, F., and Gropp, R. (2013). "Payment Defaults and Interfirm Liquidity Provision". *Review of Finance,* 1-42.

Cheng, N., and Pike, R. (2003). "The Trade Credit Decision: Evidence of UK Firms". *Managerial and decision economics*, 24, 419-438.

Deloof, M. (2003). "Does Working Capital Management Affect Profitability of Belgian Firms?". *J. of Business Finance and Accounting*, 30, 573-587.

Fawcett, T. (2006). "An Introduction to ROC Analysis". *Pattern Recognition Letters*, 27, 861-874.

García-Teruel, P., and Martínez-Solano, P. (2007). "Effects of Working Capital Management on SME Profitability". *International Journal of Managerial Finance*, 3, 174-177

Martínez-Sola, C., García-Teruel, P.J., and Martínez-Solano, P. (2012). "Trade Credit Policy and Firm Value". *IVIE Working Papers* (Serie EC).

Peel, M.J., Wilson, N., and Howorth, C. (2000). "Late Payment and Credit Management in the Small Firm Sector: Some empirical evidence". *International Small Business Journal*, 18, 17-37.

Pozuelo, J., Labatut, G., and Veres, E. (2010). "Análisis descriptivo de los procesos de fracaso empresarial en microempresas mediante técnicas multivariantes". *Revista Europea de Dirección y Economía de la Empresa*, 19, 47-66.

Seifert, D., Seifert, R.W., and Protopappa-Sieke, M. (2013). "A Review of Trade Credit Literature: Opportunities for Research in Operations". *European Journal of Operational Research*.

Terradez, M., Kizys, R., Juan, A., Debon, A., and Sawik, B. (2015). "Risk Scoring Models for Trade Credit in Small and Medium Enterprises". *Springer Proceedings in Mathematics*.

Wilner, B. (2000). "The Exploitation of Relationships in Financial Distress: The Case of Trade Credit". *The Journal of Finance*, 55, 153-178.

93. RISK ANALYSIS USING FIRST HITTING TIMES AND THRESHOLD REGRESSION: A REVIEW

X. He
University of Maryland, College Park, Department of Epidemiology and Biostatistics, USA, xinhe@umd.edu

M.-L.T. Lee
University of Maryland, College Park, Department of Epidemiology and Biostatistics, USA, mltlee@umd.edu

ABSTRACT

Statistical models for analyzing time-to-event or survival data have played an important role in risk analysis. Threshold regression (TR) based on first hitting times is a relatively new statistical approach that does not require a proportional hazard (PH) assumption in analyzing time-to-event duration data. The TR model is based on the realistic idea that a critical event occurs when an underlying stochastic process first reaches a threshold. The unknown parameters in the stochastic process and the threshold boundary are connected to covariates using suitable regression link functions. The TR model has been used to predict the probability of critical events based on duration data in many fields of application, including engineering, agriculture, business, economics, sociology, health and medicine. In the health field, for example, the TR approach has been used to model risk of lung cancer related to diesel exhaust, intervals between exacerbations in chronic obstructive pulmonary disease, risk analysis of hip fractures, among others.

1. INTRODUCTION

Statistical models for analyzing time-to-event or survival data have played an important role in risk analysis. Two widely used classes of models for time-to-event data are the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model. Other classes have also been proposed and used. Threshold regression (TR) is a relatively new model for analyzing time-to-event data.

While the Cox PH regression is a popular method for analyzing survival and time-to-event data, the proportional hazards assumption of PH regression is not always appropriate. In addition, PH regression focuses mainly on hazard ratios and thus does not offer many insights into underlying determinants of survival. Other authors have also discussed the limitations of models focussed on hazard rates, including Aalen *et al.* (2001, 2004). These limitations have led statistical researchers to explore alternative methodologies.

Threshold regression (TR) is one of these alternative methodologies. See Lee and Whitmore (2006) for a review of the first-hitting-time literature and the early development of TR models.

2. THRESHOLD REGRESSION (TR) MODEL

Previous research in several fields dealing with time-to-event data have considered models in which an event occurs when the sample path of a stochastic process of interest first satisfies a specified condition. The time until this occurrence is a first hitting time or FHT for short. The specified condition is called a boundary set or threshold. Parameters of the stochastic process and threshold can be linked to covariates using regression functions that are suited to the application at hand; hence, the name threshold regression. In mathematical terms, the FHT model has two basic components, namely, (1) a source stochastic process $\{Y(t), t \in T, y \in Y\}$ with initial value $Y(0) = y_0$, where T is the time space and Y is the state space of the process; and (2) a boundary set ß, where ß $\subset Y$. The initial value of the process y_0 is assumed to lie outside the boundary set ß. The first hitting time of ß is the random variable $T = \inf\{t : Y(t) \in ß\}$, which is the time when the stochastic process first encounters set ß. The unknown parameters in the source stochastic process $\{Y(t)\}$

and the boundary set ß may be connected to linear combinations of covariates using suitable regression link functions.

3. APPLICATIONS AND PRACTICAL EXTENSIONS

TR models and methods have been widely used. The theoretical ideas and results find immediate application in disciplines that require a study of the timing of events and duration of activities. Applications abound in health and medicine, engineering, agriculture, business, finance, economics, and the social sciences. In health and medicine, for example, recent studies have used TR to model risk of lung cancer related to diesel exhaust (Lee *et al*., 2004), recurrent exacerbations in chronic obstructive pulmonary disease (Aaron *et al*., 2015), risk analysis of hip fractures (He *et al*., 2015), among others.

Many TR applications have adopted the Wiener process to model $\{Y(t)\}$ with a variety of regression link functions. For some biological applications, see Yu *et al*. (2009), Pennell *et al*. (2009), and Lee *et al*. (2010a, 2010b). Whitmore *et al*. (1998) considered failure inference from a marker process based on a bivariate Wiener model. In engineering and science applications that involve physical degradation of a system, the degradation can only proceed in one direction so the sample path of the source stochastic process should be monotonic. A gamma process, for example, has this property and has been considered for this kind of application (see, for example, Lawless and Crowder, 2004).

In some TR applications, the natural time scale of the source stochastic process is not calendar or clock time t but rather some alternative time scale that might be called analytical time or running time. For example, wear and tear on a machine that is used intermittently might be more accurately measured by cumulative hours of operation rather than its calendar age. Denoting running time by r, its connection to calendar time involves some non-decreasing transformation r(t) where t is calendar time, with $r(0) = 0$. With this transformation, the source process $\{Y(t)\}$ of interest is defined in terms of running time r as $\{Y(r)\}$ and the subordinated process {Y [r(t)]} defines the original process in terms of calendar time t.

The connection between PH regression and TR has been examined in Lee and Whitmore (2010). They show that PH regression can be considered as a special case of

TR under suitable conditions. They give two methods of construction by which TR models can yield PH functions for survival times, one based on altering the TR time scale and the other based on varying the TR threshold or boundary set. Lee *et al*. (2010a) use an observational case study to compare empirically the benefits of TR over PH regression in understanding the influence of smoking on lung cancer. Their analysis of the study data is cross-sectional and provides no theoretical connections between TR and PH regression. The same authors, in a second investigation of the same data set, look at how TR can effectively handle longitudinal data by using a Markov decomposition procedure to break longitudinal records into a sequence of individual one-period records. Their investigation shows that PH regression is consistent with their decomposition procedure (Lee *et al*., 2010b).

He and Whitmore (2013) discussed how and to what extent AFT models are embedded within the TR family. Their discussion of the relation between TR and AFT showed that a large variety of AFT models can be derived from TR models by specifying appropriate baseline survival functions and running times.

As mentioned in Lee and Whitmore (2006), competing risks and cure rate are important extensions of the TR model. Cure rate refers to a feature of many survival models where a proportion of cases seem immune from failure. TR applications in medical research often are concerned with whether treatments can provide a cure. Lee *et al*. (2008) provides an example.

4. RECENT ADVANCES AND OPEN RESEARCH PROBLEMS

Recent advances in TR models have considered applications to longitudinal data sets with the development of associated procedures for model validation and diagnostics. Further advances are needed in this area. Multicollinearity of parameter estimates can be a practical issue. It remains to be seen which parameterizations of threshold regression models tend to have relatively independent estimation errors. Multicollinearity within regression functions will tend to show itself in familiar ways and will likely be dealt with by conventional remedies. Variable selection is another topic to be considered. Nonparametric, semiparametric and other robust estimation methods seem to have much to contribute to the successful application of threshold

regression. Quasi-likelihood methods and generalized estimating equations may offer feasible approaches. Applications of TR in settings where data are gathered from complex sample surveys is another open area of research. Much research remains to be done on threshold regression and many new tools are needed. These tools are likely to be developed as threshold regression is applied to a broader array of practical settings and new situations are encountered.

REFERENCES

Aalen, O.O., and Gjessing, H.K. (2001). "Understanding the Shape of the Hazard Rate: A Process Point of View". *Statistical Science*, 16, 1-22.

Aalen, O.O., and Gjessing, H.K. (2004). "Survival Models Based on the Ornstein-Uhlenbeck Process". *Lifetime Data Analysis*, 10, 407-423.

Aaron, S.D., Stephenson, A.L., Cameron, D.W., and Whitmore, G.A. (2015). "A Statistical Model to Predict One-Year Risk of Death in Patients with Cystic Fibrosis". *Journal of Clinical Epidemiology*, in press.

He, X., and Whitmore, G.A. (2013). "Connecting threshold regression and accelerated failure time models". Lee, M.-L.T., Gail, M., Pfeiffer, R., Satten, G., Cai, T., and Gandy, A. (eds.), *Risk Assessment and Evaluation of Predictions*. New York: Springer, pp. 47-60.

He, X., Whitmore, G.A., Loo, G.Y., Hochberg, M.C., and Lee, M.-L.T. (2015). "A Model for Time to Fracture with a Shock Stream Superimposed on Progressive Degradation: The Study of Osteoporotic Fractures". *Statistics in Medicine*, 34, 652-663.

Lawless, J., and Crowder, M. (2004). "Covariates and Random Effects in a Gamma Process Model with Application to Degradation and Failure". *Lifetime Data Analysis*, 10, 213-227.

Lee, M.-L.T., and Whitmore, G.A. (1993). "Stochastic Processes Directed by Randomized Time". *Journal of Applied Probability*, 30, 302-314.

Lee, M.-L.T., Garshick, E., Whitmore, G.A., Laden, F., and Hart, J. (2004). "Assessing Lung Cancer Risk to Rail Workers Using a First Hitting Time Regression Model". *Environmetrics*, 15, 1-12.

Lee, M.-L.T., and Whitmore, G.A. (2006). "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary". *Statistical Sciences*, 21, 501-513.

Lee, M.-L.T., Chang, M., and Whitmore, G.A. (2008). "A Threshold Regression Mixture Model for Assessing Treatment Efficacy in a Multiple Myeloma Clinical Trial". *Journal of Biopharmaceutical Statistics*, 18, 1136-1149.

Lee, M.-L.T., and Whitmore, G.A. (2010). "Proportional Hazards and Threshold Regression: Their Theoretical and Practical Connections". *Lifetime Data Analysis*, 16, 196-214.

Lee, M.-L.T., Whitmore, G.A., and Rosner, B. (2010a). "Benefits of Threshold Regression: A Case-Study Comparison with Cox Proportional Hazards Regression". In: Rykov, V.V., Balakrishnan, N., Nikulin, M.S. (eds.), *Mathematical and Statistical Models and Methods in Reliability*, Boston: Birkhauser, pp 359-370.

Lee, M.-L.T., Whitmore, G.A., and Rosner, B. (2010b). "Threshold Regression for Survival Data with Time-Varying Covariates". *Statistics in Medicine*, 29, 896-905.

Pennell, M.L., Whitmore, G.A., and Lee, M.-L.T. (2009). "Bayesian Random Effects Threshold Regression with Application to Survival Data with Non-Proportional Hazards". *Biostatistics*, 11, 111-126.

Whitmore, G.A., Crowder, M.J., and Lawless, J.F. (1998). "Failure Inference from a Marker Process Based on a Bivariate Wiener Model". *Lifetime Data Analysis*, 4, 229-251.

Yu, Z., Tu, W., and Lee, M.-L.T. (2009). "A Semiparametric Threshold Regression Analysis of Sexually Transmitted Infections in Adolescent Women". *Statistics in Medicine*, 28, 3029-3042.

94. STOCHASTIC BOUNDS FOR TIMES ELAPSED BETWEEN SUCCESSIVE FAILURES OF COMPONENTS IN RELIABILITY SYSTEMS

N. Torrado
University of Coimbra, Centre for Mathematics (CMUC), Portugal,
nuria.torrado@gmail.com

R. E. Lillo
Universidad Carlos III de Madrid, Department of Statistics, Spain,
rosaelvira.lillo@uc3m.es

ABSTRACT

In the reliability context, the times elapsed between successive failures of components in a system are called spacings. Spacings are of great interest in other many areas. In particular, in auction theory, the second and the last spacings represent reverse auction in the second-price business auction and auction rent's in buyer's auction, respectively. In addition, there are many goodness-of-fit test based on spacings very useful in the context of life testing.

In the article and in the presentation, we investigate stochastic relations in terms of stochastic orderings between spacings of a random sample from independent but not identically distributed random variables. In particular, we solve some open problems for spacings from a sample and also between spacings from two samples, in both cases we consider that the observations are heterogeneous but independent exponentially distributed.

1. INTRODUCTION

If the random variables $X_1, \ldots, X_n$ are arranged in ascending order of magnitude, then the i'th smallest of X_i's is denoted by $X_{i:n}$. The ordered quantities

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n},$$

are called order statistics, and $X_{i:n}$is the i'th order statistic. These random variables are of great interest in many areas of statistics, in particular, in the characterizations of distributions and testing problems, among others. Specifically, there is a very interesting application of OS's in reliability theory. The $(n-k+1)$'th OS in a sample of size n represents the life length of a k-out-of-n system which is an important technical structure. It consists of n components of the same kind with independent and identically distributed life lengths. All n components start working simultaneously, and the system works, if at least k components function; i.e. the system fails, if $(n-k+1)$ or more components fail.

Here, we are interested in another random variables which correspond to times elapsed between successive failures in the reliability context and are called simple spacings. The i'th simple spacing is defined as

$$D_{i:n} = X_{i:n} - X_{i-1:n},$$

and the i'th normalized spacing is defined as

$$D^*_{1:n}, \ldots, D^*_{n,n}$$

Many authors have studied the stochastic properties of spacings from independent and identically distributed (iid) random variables. In this context, the exponential distribution possesses a lack of ageing property, so that, the failure rate is constant and the spacings correspond to times elapsed between successive failures of components in a system. The following remarkable property of the exponential distribution was first demonstrated by Sukhatme (1937).

Theorem 1.1. If $X_1, X_2, \ldots, X_n$ are independent and identically exponential random variables, then $D^*_{1:n}, \ldots, D^*_{n,n}$ are iid random variables having the same exponential distribution.

Such a characterization may not hold for other distributions or when the observations are not independent and identically distributed. In the conventional modelling of these structures, the component lifetimes are supposed to be independent and identically distributed random variables. However, in many practical situations, like in reliability theory, the observations are not necessarily iid. The objective of this work is first to discuss some recent results on stochastic comparisons between spacings of heterogeneous samples and present some extensions. Most of the results presented here can be found in Torrado and Lillo (2014). Further results on this topic can be found, for instance, in Kochar and Kowar (1996), Kochar and Rojo (1996), Wen *et al*. (2007), Hu *et al*. (2008), Kochar and Xu (2011) and Torrado and Lillo (2013).

The article is organized as follows. In Section 2 we recall some definitions and we present our main results in Section 3.

2. DEFINITIONS

In this section, we give briefly a review of some definitions. First, Recall that the Glaser's function ηX of a random variable X (see Glaser (1980)), at the point t is defined as

$$\eta_X(t) = -\frac{f'(t)}{f(t)} = -\big(\log f(t)\big)',$$

where f is the density function of X.

Definition 2.1. Let η_Y be the Glaser's function of another random variable Y. We say that X is smaller than Y in the likelihood ratio order if $\eta_X(t) \geq \eta_Y(t)$ for all t, denoted by $X \leq_{lr} Y$.

Next, let us introduce some notations and definitions. Let $x = (x_1, x_2, \ldots, x_n)$, $n > 1$ a vector with positive components, then

$$S(\mathbf{x}) = \sum_{i=1}^{n} x_i\,, \tag{1}$$

is the first elementary symmetrical function of the positive $x_1, x_2, \ldots, x_n$. Following Torrado *et al*. (2010) the density function of D_i can be written as

$$f_i(t) = \sum_{j=1}^{M_i} \Delta(\beta_{m_j}^{(i)}, n)\beta_{m_j}^{(i)} e^{-t\beta_{m_j}^{(i)}}, \tag{2}$$

with $M_i = \binom{n}{n-i+1}$,

$$\beta_{m_j}^{(i)} = S(\lambda_{m_j}),$$

where $\mathrm{S}(\cdot)$ is defined as in (1) and m_j indicates a group of indices of size $n - \mathrm{i} + 1$, and

$$\Delta(\beta_{m_j}^{(i)}, n) = \sum_{\mathbf{r}_{i-1,m_j}} \left(\prod_{k \in H_{m_j}} \lambda_k \right) \left[\prod_{\ell=1}^{i-1} \left\{ \sum_{\substack{u=\ell \\ r_u \in H_{m_j}}}^{i-1} \lambda_{r_u} + \beta_{m_j}^{(i)} \right\} \right]^{-1},$$

where $H_{mj} = \{1, \ldots, n\}$ – mj and the outer summation is being taken over all permutations of the elements of H_{mj}.

Observing eq. (2), note that $D_{i:n} \leq_{lr} D_{i+1:n}$ if and only if

$$\sum_{j=1}^{M_{i+1}} \sum_{k=1}^{M_i} \Delta(\beta_{m_k}^{(i)}, n)\Delta(\beta_{m_j}^{(i+1)}, n)\beta_{m_k}^{(i)}\beta_{m_j}^{(i+1)} e^{-t\left(\beta_{m_k}^{(i)} + \beta_{m_j}^{(i+1)}\right)} \left(\beta_{m_k}^{(i)} - \beta_{m_j}^{(i+1)}\right) \geq 0.$$

Throughout this paper we suppose without loss of generality that the λ_i's are in increasing order.

3. MAIN RESULTS

In this section, we study conditions under which simple spacings from a sample and also spacings from two samples are ordered in the likelihood ratio ordering.

3.1. A sample problem

Here we consider a sequence of random variables, $X_1, \ldots, X_n$, of independent exponential random variables with X_i having hazard rate λ_i, for $i = 1, \ldots, n$.

In the following results, we prove that the second spacing is always smaller than the third spacing in terms of the likelihood ratio order and we provide the ordering among all spacings in the case $n = 4$ (see Torrado and Lillo (2014) for details).

Theorem 3.1. (see Torrado and Lillo (2014)) Let $X_1, \ldots, X_n$ be independent exponential random variables such that X_i has hazard rate λ_i for $i = 1, \ldots, n$, then

$$D_{2:n} \leq_{lr} D_{3:n}.$$

Hu *et al.* (2008) proved

$$D_{1:3} \leq_{lr} D_{2:3} \leq_{lr} D_{3:3}$$

In the following result, we extend the result in Hu *et al.* (2008) to $n = 4$. Hu *et al.* (2008) also proved that $D_{1:n} \leq_{lr} D_{2:n}$ for any n, and by Theorem 3.1 we know that $D_{2:n} \leq_{lr} D_{3:n}$ for any n, so we have to show that $D_{3:4} \leq_{lr} D_{4:4}$.

Theorem 3.2. (see Torrado and Lillo (2014)) Let $X_1, \ldots, X_4$ be independent exponential random variables such that X_i has hazard rate λ_i for $i = 1, \ldots, 4$, then

$$D_{3:4} \leq_{lr} D_{4:4}$$

3.2. Two samples problem

In the two samples problem, let $X_1, \ldots, X_n$ be independent exponential random variables with X_i having hazard rate λ_i, $i = 1, \ldots, n$ and $Y_1, \ldots, Y_n$ another random sample with Y_i having hazard rate θ_i, $i = 1, \ldots, n$. For $\lambda_1 \leq \lambda_2 \leq \lambda_*$, Hu *et al.* (2006) showed that spacings are increasing in the likelihood ratio order for $n = 2, 3$. In the following result, we prove this open problem for the second spacing for arbitrary n.

Theorem 3.3. (see Torrado and Lillo (2014)) Let $X_1, \ldots, X_n$ be independent exponential distributions such that X_i has hazard rate λ_1, for $i = 1, \ldots, p$ and X_j has hazard rate λ_* for $j = p + 1, \ldots, n$. Let $Y_1, \ldots, Y_n$ be another set of independent exponential distributions such that Yi has hazard rate $\lambda 2$, for $i = 1, \ldots, p$ and Y_j has hazard rate λ_* for $j = p + 1, \ldots, n$. If $\lambda_1 \leq \lambda_2 \leq \lambda_*$, then

$$C_{2:n} \leq_{lr} D_{2:n}$$

where $D_{2:n}$ and $C_{2:n}$ are the second simple spacing from X_i's and Y_i's, respectively.

Recall that normalized spacings are defined as $D^*_{i:n} = (n - i + 1)D_{i:n}$, for i = 1, ... , n. Note that, from Theorem 3.1 in Torrado and Lillo (2013) and Theorem 3.3, we have that

$$\lambda_1 \leq \lambda_2 \leq \lambda_* \Rightarrow C^*_{2:2} \leq_{lr} D^*_{2:2},$$

where $D^*_{2:2}$ and $C^*_{2:2}$ are the second normalized spacings from two multiple-outlier exponential models as before.

ACKNOWLEDGMENTS

This work was supported by the Portuguese Government through the Fundação para a Ciência e a Tecnologia under the grant SFRH/BPD/91832/2012 and partially by project of the Spanish Ministry of Education and Science (ECO2012-38442).

REFERENCES

Glaser, R.E. (1980). "Bathtub and Related Failure Rate Characterizations". *Journal of the American Statistical Association*, 75, 667-672.

Hu, T., Wang, F., and Zhu, Z. (2006). "Stochastic Comparisons and Dependence of Spacings from two Samples of Exponential Random Variables". *Communications in Statistics - Theory and Methods*, 35, 979-988.

Hu, T., Lu, Q., and Wen, S. (2008). "Some New Results on Likelihood Ratio Orderings for Heterogeneous Exponential Random Variables". *Communications in Statistics - Theory and Methods*, 37, 2506-2515.

Kochar, S.C., and Korwar, R. (1996). "Stochastic Orders for Spacings of Heterogeneous Exponential Random Variables". *J. Multivariate Anal.*, 57, 69-83.

Kochar, S.C., and Rojo, J. (1996). "Some New Results on Stochastic Comparisons of Spacings from Heterogeneous Exponential Distributions". *J. Multivariate Anal.*, 59, 272-281.

Kochar, S.C., and Xu, M. (2011). "Stochastic Comparisons of Spacings from Heterogeneous Samples". In Rao, J.S., Wells, M., and Sengupta, A. (eds.), *Advances in Directional and Linear Statistics: A Festschrift*, Springer, New York, pp. 113-129.

Sukhatme, P.V. (1937). "Tests of Significance for Samples of the C2 Population with two Degrees of Freedom". *Annals of Eugenics*, 8, 52-56.

Torrado, N., Lillo, R.E., and Wiper, M.P. (2010). "On the Conjecture of Kochar and Korwar". *J. Multivariate Anal.*, 101, 1274-1283.

Torrado, N., and Lillo, R.E. (2013). "Likelihood ratio order of spacings from two heterogeneous samples". *J. Multivariate Anal.,* 114, 338-348.

Torrado, N., and Lillo, R.E. (2014). "Likelihood Ratio Comparisons among Spacings Related to both one or two Samples". To appear in *Statistics*. DOI: 10.1080/02331888.2014.908192.

Wen, S., Lu, Q., and Hu, T. (2007). "Likelihood Ratio Orderings of Spacings of Heterogeneous Exponential Random Variables". *J. Multivariate Anal.*, 98, 743-756.

95. MULTIVARIATE RISK MEASURES: A DIRECTIONAL APPROACH FORA VALUE AT RISK

R. Torres
Universidad Carlos III de Madrid, Statistics Department, Madrid, ratorres@est-econ.uc3m.es

R. E. Lillo
Universidad Carlos III de Madrid, Statistics Department, Madrid, rosaelvira.lillo@uc3m.es

H. Laniado
Universidad Carlos III de Madrid, Statistics Department, Madrid, hlaniado@est-econ.uc3m.es

ABSTRACT

Value at risk (VaR) is a widely used measure in risk management defined as a quantile associated to the loss distribution. VaR has both good analytic properties and easy interpretation. However, its extension to the multivariate framework is not unique because a unique definition of multivariate quantile does not exist. In the current literature, the multivariate quantiles are related to a specific partial order considered in $\mathbb{R}^n$, or to a property of the univariate quantile that is desirable to be extended to $\mathbb{R}^n$.

In this work, we introduce a multivariate value at risk as a vector-valued directional risk measure, based on a directional multivariate quantile introduced recently in the literature. The directional approach allows the manager to consider external information or risk preferences in her/his analysis. We have proven some interesting properties of this risk measure.

1. INTRODUCTION

Value at risk (VaR) has become a benchmark for risk management. For univariate risks It is defined as the α–quantile of the loss distribution function. Thus, the VaR is a measure easily interpretable, and it still remains as the most popular measure used by risk managers which it is commonly implemented by investment banks to measure the market risk of their asset portfolios. Although (VaR) has been broadly criticized from the work of Artzner *et al.* (1999) since it does not verify the diversification property, it has also been defended by Heyde *et al.* (2009) for its robustness.

Unfortunately, a unique definition of multivariate VaR is more complicated because there are different possible definitions of multidimensional quantiles trying to generalize some desirable properties of the univariate quantile. For instance, the proposals given by Koltchinskii (1997) of multivariate quantiles as inversions of mappings, multivariate quantiles in terms based on norm minimization as in Chaudhuri (1996), multivariate quantiles as level-sets given by Fernández-Ponce and Suárez-Llorens (2002), multivariate quantiles based on depth functions developed in Serfling (2002), and finally, multivariate quantiles based on projections as in Fraiman and Pateiro-López (2012) and Kong and Mizera (2012).

Currently business and financial activities generate data for which it has been shown that it is insufficient to consider single real-value measures over marginal aspects. For instance, during the 2007-2009 crisis, weakness in the solvency and liabilities dependence among the financial institution branches have been detected. Therefore, the regulations have been changed to complement the risk management using internal models in the institutions that have to include multivariate risk measures computable in high dimensions and also, to consider possible internal and external risks to obtain better hedge results, even if the nature of those risks is strongly heterogeneous.

In recent decades, literature devoted to extend the VaR measure to the multivariate setting has been published. For instance, the notions of VaR introduced by Lee and Prékopa (2012), Embrechts and Puccetti (2006) and Cousin and Di Bernardino (2013). We extend those works in this paper by introducing a directional multivariate Value at Risk, based on the extremality level sets introduced in Laniado *et al.* (2012). The

risk measure that we propose considers both the high dimension nature of the real problems and the dependence among the risks implied in the analysis. Finally, we give the possibility of considering manager preferences, introducing a parameter of direction u. For instance, directions like the maximum variability given for the principal components in the portfolio, or the assets weight composition could be more interesting to analyze than the classic directions given for the information summarized in the survival or cumulative distribution functions. Besides, the directional approach allows us to give bounds for the VaR related to linear combination of random variables, mainly when they are statistically dependent.

We have proved properties of the directional VaR that we consider as relevant for a multivariate risk measure, such as consistency with respect to a particular stochastic order and tail subadditivity in the mean loss direction, as well as some invariance properties. In Section 2, we introduce some preliminary concepts and notation necessary in order to understand the main contributions of the paper. The main contributions are given in Section 3.

2. PRELIMINARY CONCEPTS AND NOTATION

We have devoted this section to revise the main concepts that are necessary to properly define the risk measure introduced in this paper.

Definition 2.1. An oriented orthant in $\mathbb{R}^n$ with vertex x in the direction u is defined as,

$$\mathfrak{C}_{\mathbf{x}}^{\mathbf{u}} = \{\mathbf{z} \in \mathbb{R}^n : R_u(\mathbf{z} - \mathbf{x}) \geq 0\}, \tag{1}$$

where $\mathbf{u} \in \bar{\mathbb{B}}_n(0) = \{\mathbf{v} \in \mathbb{R}^n : ||\mathbf{v}|| = 1\}$ and R_{u} is the orthogonal matrix such that $R_{\mathrm{u}}\mathbf{u} = e$, with $\mathbf{e} = \frac{\sqrt{n}}{n}[1, ..., 1]'$.

Based on the oriented orthant concept, we can define a partial data order (denoted by $\preceq_{\mathbf{u}}$) in $\mathbb{R}^n$ as,

$$\mathbf{x} \preceq_{\mathbf{u}} \mathbf{y}, \text{ if and only if, } \mathfrak{C}_{\mathbf{x}}^{\mathbf{u}} \supseteq \mathfrak{C}_{\mathbf{y}}^{\mathbf{u}}, \tag{2}$$

where $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$. Or equivalently,

$$\mathbf{x} \preceq_{\mathbf{u}} \mathbf{y}, \text{ if and only if, } R_{\mathbf{u}}\mathbf{x} \leq R_{\mathbf{u}}\mathbf{y},$$

where the order on the right side is component-wise.

Throughout the paper we will use the following notation related to subsets in $\mathbb{R}^n$. Given $\mathrm{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, and $A \subset \mathbb{R}^n$, the sets $\mathrm{b} + A$ and cA are defined as,

$$\mathbf{b} + A := \{\mathbf{b} + \mathbf{a} : \mathbf{a} \in A\}, \qquad cA := \{c\mathbf{a} : \mathbf{a} \in A\}. \tag{3}$$

The following stochastic order defined in Laniado *et al.* (2012) will be a key tool in providing some properties of the multivariate VaR that we will define in the next Section.

Definition 2.2. Let X and Y be two random vectors in $\mathbb{R}^n$, X is said smaller than Y in the extremality order in the direction u (denoted by $\mathrm{X} \leq \mathcal{E}_{\mathrm{u}}\, \mathrm{Y}$) if,

$$\mathbb{P}\left[R_{\mathbf{u}}(\mathbf{X} - \mathbf{z}) \geq 0\right] \leq \mathbb{P}\left[R_{\mathbf{u}}(\mathbf{Y} - \mathbf{z}) \geq 0\right], \text{ for all } z \text{ in } \mathbb{R}^n.$$

It is easy to show that $\mathrm{X} \leq \mathcal{E}_{\mathrm{u}}\, \mathrm{Y} \leftrightarrow R_{\mathrm{u}}\mathrm{X} \leq_{uo} R_{\mathrm{u}}\mathrm{Y}$. Moreover, if $\mathrm{X} \leq \mathcal{E}_{\mathrm{u}}\, \mathrm{Y}$ then $\mathbb{E}[\mathbf{X}] \preceq_{\mathbf{u}} \mathbb{E}[\mathbf{Y}]$, as it is proven in Laniado *et al.* (2012), Property 3.4. Since the multivariate VaR is based on the definition of a quantile, we also need to introduce the directional multivariate quantile given in Laniado *et al.* (2010).

Definition 2.3. Let X be a random vector with associated probability distribution function $\mathbb{P}$. Then the directional multivariate quantile at level α, in direction u is defined as

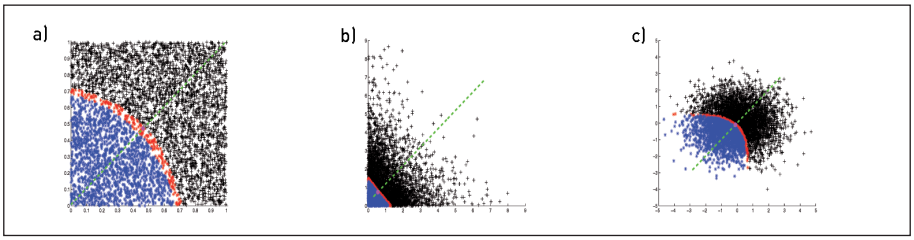
$$\mathcal{Q}_{\mathbf{X}}(\alpha, \mathbf{u}) := \partial\{\mathbf{x} \in \mathbb{R}^n : \mathbb{P}(\mathfrak{C}_{\mathbf{x}}^{\mathbf{u}}) \leq \alpha\}, \tag{4}$$

with $0 \leq \alpha \leq 1$.

From now on, we will focus on an absolutely-continuous random vector X (with respect to the Lebesgue measure ν on $\mathbb{R}^n$) with increasing marginal distribution functions and such that $\mathbb{E}[X_i] < \infty$, for $i = 1, \ldots, n$. These conditions will be called regularity conditions.

3. DIRECTIONAL VALUE AT RISK AND PROPERTIES

We propose a definition of multivariate VaR for a vector of n-dependent risks linked with the directional multivariate quantile defined in (4). Besides,



(a) Bivariate Uniform, (b) Bivariate Exponential, (c) Bivariate Normal.
Figure 1. $VaR^{-\mathbf{e}}_{0.7}(\mathbf{X})$.

the output is a point in $\mathbb{R}^n$; that is, a vector of the same dimension as the considered vector of risks.

Definition 3.1. Let X be a random vector satisfying the regularity conditions and $0 \leq \alpha \leq 1$. Then the directional multivariate Value at Risk of X in direction u at probability level α is given by

$$VaR^{\mathbf{u}}_{\alpha}(\mathbf{X}) = \left(\mathcal{Q}_{\mathbf{X}}(\alpha, \mathbf{u}) \bigcap \{\lambda \mathbf{u} + \mathbb{E}[\mathbf{X}]\} \right), \tag{5}$$

where $\lambda \in \mathbb{R}$.

As you can see, given a direction u, the $VaR^{\mathbf{u}}_{\alpha}(\mathbf{X})$ is the intersection between the directional quantile at level α, and the line defined by both the direction u and the mean of X. We have shown the choice of the mean in the definition of (5) allows us to derive desirable and interpretable analytic properties related to the risk measure. However, other options as central reference point are possible; for example the median seen as the deepest point associated with a multivariate depth measure, which may provide a more robust risk measure (e.g. Zuo and Serfling, 2000).

To illustrate this concept, some examples of the risk measure defined in (5) are given in Figure 1 for three different bivariate distributions in the direction e and for $\alpha = 0.3$. If

you consider any point on the blue region as a vertex of an oriented orthant in direction $\mathbf{e}$, then the probability of that orthant will be greater than α. It will be equal to α or smaller than α if the point is taken from the red line or black region, respectively.

We have obtained some properties of $VaR_{\alpha}^{\mathbf{u}}(\mathbf{X})$ that are similar to those properties considered in the risk literature (see Artzner *et al.,* 1999), such as properties related to coherent risk measures, but related to the dimension n. Besides, we have explored other properties inherent to the multivariate response such as invariance under orthogonal transformations.

ACKNOWLEDGMENTS

This research was partially supported by a Spanish Ministry of Economyand Competition grant ECO2012-38442.

REFERENCES

Arbia, G. (2002). "Bivariate Value-at-Risk". *Statistica* LXII, 231-247.

Artzner, P., Delbae, F., Heath, J., and Eber, D. (1999). "Coherent Measures of Risk". *Mathematical Finance* 3, 203-228.

Cousin, A., and Di Bernardino, E. (2013). "On Multivariate Extensions of Value-at-Risk". *Journal of Multivariate Analysis* 119, 32-46.

Danielson, J., Jorgensen, B., Samorodnitsky, G., Sarma, M., and De Vries, C. (2013). "Fat Tails, VaR and Subadditivity". *Journal of econometrics* 172 (2), 283-291.

Embrechts, P., and Puccetti, G., (2006). "Bounds for Functions of Multivariate Risks". *Journal of Multivariate Analysis* 97, 526-547.

Fernández-Ponce, J., and Suárez-Llorens, A. (2002). "Central Regions for Bivariate Distributions". *Austrian Journal of Statistics* 31, 141-156.

Fraiman, R., and Pateiro-López, B. (2012). "Quantiles for Finite and Infinite Dimensional Data". *Journal Multivariate Analysis* 108, 1-14.

Heyde, C., Kou, S., and Peng, X. (2009). "What Is a Good External Risk Measure: Bridging the Gaps between Robustness, Subadditivity, and Insurance Risk Measure" (preprint).

Jessen, A., and Mikosh, T. (2006). "Regularly Varying Functions". *Publications de l'Institute Mathématique* 79.

Koltchinskii, V. (1997). "M-Estimation, Convexity and Quantiles". *Ann. Statist.* 25 (2), 435-477.

Kong, L., and Mizera, I. (2012). "Quantile Tomography: Using Quantiles with Multivariate Data". *Statistica Sinica* 22 (4), 1589-1610.

Lee, J., and Prékopa, A. (2012). "Properties and Calculations of Multivariate Risk Measures: MVaR and MCVaR". *Rutcor Research Report* 25.

Laniado, H., Lillo, R., Pellerey, F., and Romo, J. (2012). "Portfolio Selection through an Extremality Stochastic Order". *Insurance: Mathematics and Economics*, 51, 1-9.

Laniado, H., Lillo, R., and Romo, J. (2010). "Multivariate Extremality Measure". Working Paper, *Statistics and Econometrics*, Series 08, Universidad Carlos III de Madrid.

Nappo, G., and Spizzichino, F. (2009). "Kendall Distributions and Level Sets in Bivariate Exchangeable Survival Models". *Information Sciences* 179, 2878-2890.

Resnick, S. (1987). *Extreme values, regular variation and point process.* Springer-Verlag.

Serfling, R. (2002). "Generalized Quantile Processes Based on Multivariate Depth Function, with Applications in Nonparametric Multivariate Analysis". *Journal of Multivariate Analysis* 83, 232-247.

Zuo, Y., and Serfling, R. (2000). "General Notions of Statistical Depth Function". *Annals of Statistics* 28 (2), 461-482.

96. LOCAL MAXIMUM BASED TAIL INDEX ESTIMATOR

M. Vaičiulis
Vilnius university, Institute of Mathematics and Informatics,Lithuania,
marijus.vaiciulis@mii.vu.lt

ABSTRACT

We continue the investigation of the tail index estimators, which are based on maxima over blocks of data and which were proposed in Paulauskas (2003), Paulauskas and Vaičiulis (2011) and Paulauskas and Vaičiulis (2012). We construct a new tail index estimator, based on local maxima, and we prove the weak consistency and asymptotic normality of this new estimator in the case of i.i.d. observations. Also the weak consistency is proved in the case where data is generated by a non-stationary max-MA(q) process.

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (i.i.d.) and nonnegative random variables with a common distribution function F. We suppose that the tail $1-F$ is regularly varying with the index $1/\gamma > 0$, i.e.,

$$1 - F(x) = h(x)x^{-1/\gamma}, \quad x \to \infty, \tag{1}$$

where *h(x)* is a slowly varying at infinity function. In the statistical literature a parameter γ (in general, $\gamma \in \mathbf{R}$) is referred to as the extreme value index, while a parameter $\alpha = 1/\gamma$ is a tail index.

The majority of the existing estimators of the tail index α are constructed by using the order statistics $X_{n,1} \leq \cdots \leq X_{n,n}$ of the observations $X_1, \ldots, X_n$, see e.g. review Gomes *et al*. (2008).

Other (or alternative) approach of the tail index α estimation uses grouping of the observations $X_1, X_2, \ldots, X_n$ into groups $Gk = \{X_{(k-1)m+1}, \ldots, X_{km}\}$, $k = 1, 2, \ldots, [n/m]$, where $2 \le m \le n$ and $[\cdot]$ denotes the integer part.

Several such kind estimators for the tail index α can be written in terms of so-called generalized Davydov, Paulauskas and Račkauskas (further - DPR) statistic, which was provided in Paulauskas and Vaičiulis(2011) and has form

$$S_{n,m}(f) = \frac{1}{[n/m]} \sum_{k=1}^{[n/m]} f\left(\frac{\ell_{m,k}}{L_{m,k}}\right), \tag{2}$$

where $L_{m,k}$ and $\ell_{m,k}$ denote the the largest and the second largest element in the group Gk (containing m elements) and a function $f: [0, 1] \to (-\infty, +\infty)$ is such that $\mathrm{E}f(1/Y) < \infty$, where Y is a standard Pareto random variable. Taking $f = f^{(1)}, f^{(1)}(t) = t$ in (2) we obtain the so-called DPR estimator (proposed in Paulauskas, 2003; see also Davydov and Paulauskas, 1999) for a parameter $\alpha/(\alpha+1)$, while the DPR estimator for the parameter α has form

$$\hat{\alpha}_n^{(1)}(m) = \frac{S_{n,m}(f^{(1)})}{1 + S_{n,m}(f^{(1)})}.$$

A partial case of Qi (2010) estimator for the tail index can be presented as

$$\hat{\alpha}_n^{(2)}(m) = \frac{1}{S_{n,m}(f^{(2)})},$$

where $f^{(2)}(t) = ln(1/t)$. Paulauskas and Vaičiulis (2011) introduced the tail index estimator with tuning parameter r:

$$\hat{\alpha}_n^{(3)}(m, r) = \frac{1 - rS_{n,m}(f_r^{(3)})}{S_{n,m}(f_r^{(3)})},$$

where $f^{(3)}(t) = (1-t^r)/r$ for $r \neq 0$ and $f_r^{(3)}(0) = \lim_{r\to 0} f_r^{(3)}(t)$. This estimator is a generalization of the estimators $\hat{\alpha}_n^{(1)}(m)$ and $\hat{\alpha}_n^{(2)}(m)$, since $\hat{\alpha}_n^{(3)}(m, 1) = \hat{\alpha}_n^{(1)}(m)$ and $\hat{\alpha}_n^{(3)}(m, 0) = \hat{\alpha}_n^{(2)}(m)$.

Under stronger than (1) assumption the asymptotic behavior of the estimator $\hat{\alpha}_n^{(3)}(m, r)$ in max-aggregation context was investigated in Paulauskas and Vaičiulis (2012).

Theorem 1.1. Let $X_1, \ldots, X_n$ be a sample obtained by using scheme

$$X_t = \max\{\xi_{t-q}, \ldots, \xi_t\}, \quad t \in \mathbf{Z}, \tag{3}$$

where $q \geq 1$ is a fixed integer and $\xi t, t \in Z$ are independent and identically distributed random variables with a distribution function F satisfying $1 - F(x) \sim C x^{-\alpha}$, $x \to \infty$, where $C > 0$ and $\alpha > 0$. Suppose that $-\alpha/2 < r < \infty$, and let us suppose that a sequence $m = m(n)$ is such that $m(n) \to \infty$ and $n/m(n) \to \infty$ as $n \to \infty$. Then

$$S_{n,m}(f_r^{(3)}) \xrightarrow{\mathrm{P}} \frac{1}{r+\alpha}, \quad \hat{\alpha}_n^{(3)}(m,r) \xrightarrow{\mathrm{P}} \alpha, \tag{4}$$

as $n \to \infty$. Here and below $\xrightarrow{\mathrm{P}}$ denotes convergence in probability.

Unfortunately, the last result can not be extended for a general max-MA(q) process

$$X_t = \xi_t \vee (C_1 \xi_{t-1}) \vee (C_q \xi_{t-q}), \quad t \in \mathbf{Z}, \tag{5}$$

where $C_j \geq 0$, $1 \leq j \leq q$. Let C^* and C_* denote the largest and the second largest of the weight $1, C_1, \ldots, C_q$. Following the lines in Paulauskas and Vaičiulis (2012) one can prove that in the case $\mathrm{C}^* \neq \mathrm{C}_*$ statistic $S_{n,m}(f_r^{(3)})$ tends to $(\alpha + r)^{-1} (1 - (\mathrm{C}^*/\mathrm{C}_*)^{\alpha+r})$ in probability. Moreover, under several assumptions on a function f in (2) (see Vaičiulis, 2012), the following result holds: if $\mathrm{C}^* \neq \mathrm{C}_*$, then generalized DPR statistic $S_{n,m}(f)$ tends to $f(1) - \int_{C_*/C^*}^1 t^\alpha \mathrm{d}f(t)$ in probability. Thus, if $X_1, \ldots, X_n$ is a sample from a process (5), then $\hat{\alpha}_n^{(1)}(m)$, $\hat{\alpha}_n^{(2)}(m)$ and $\hat{\alpha}_n^{(3)}(m,r)$ are not consistent estima- n n n tors for the parameter $\alpha > 0$. Detail investigation of the bias of the statistic $S_{n,m}(f)$ allows to deduce that removing the second largest term $\ell_{m,k}$ from the construction of (2) solve the consistency problem.

Put $g(x, y) := \ln((x \vee y)/(x \wedge y))$. We introduce the following estimator of the parameter $\gamma > 0$, based on the local maximums

$$\hat{\gamma}_n(m) = \frac{1}{\tilde{m} \ln(4)} \sum_{k=1}^{\tilde{m}} g\left(L_{m,k}, L_{m,k+1}\right),$$

where parameter m satisfies $1 \leq m \leq [n/2]$ and $\tilde{m} := [n/m] - 1$. Construction of this estimator is based on the identity

$$\mathrm{E} g\left(\frac{\max\{Y_1, \ldots, Y_m\}}{m^\gamma}, \frac{\max\{Y_{m+1}, \ldots, Y_{2m}\}}{m^\gamma}\right) = \gamma \ln \tag{4},$$

where $Y_1, \ldots, Y_{2m}$ are independent Fréchet random variables with the distribution function $G_\alpha(x) = \exp\{-x^{-1/\gamma}\}$, $x > 0$. Note that the estimator $\hat{\gamma}_n(m)$, as well as the estimator $\hat{\alpha}_n(m) = 1/\hat{\gamma}_n(m)$ for the tail index $\alpha > 0$, is scale-free, i.e., it does not change when X_k is replaced by cX_k with $c > 0$.

2. ASYMPTOTIC RESULTS

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed and non-negative random variables with a common distribution function F. Denote $-1/\ln(F))^{\leftarrow}(t)$, $t \geq 1$, where $W^{\leftarrow} : I \to \mathbb{R}$ is the left continuous inverse function of a monotone function W, defined by the relation $W^{\leftarrow}(t) = \sup\{x : W(x) \leq t\}$, $t \in I$. We suppose that the function V is regularly varying with the index $\gamma > 0$ (written $V \in RV_\gamma$), i.e., for $x > 0$,

$$\lim_{t\to\infty} \{\ln(V(tx)) - \ln(V(t))\} = \gamma \ln(x). \tag{6}$$

Limit (6) is equivalent to (1), see De Haan and Ferreira (2006).

Theorem 2.1. Let $X_1, \ldots, X_n$ be i.i.d. random variables with a distribution function F satisfying $F(1) = 0$ and (6). Let us suppose that given a sequence $m \equiv m(n)$ satisfying

$$m(n) \to \infty, \quad \frac{n}{m(n)} \to \infty, \quad n \to \infty. \tag{7}$$

Then $\hat{\gamma}_n(m)$ is weakly consistent estimator for the parameter γ, i.e., the sequence $\{\hat{\gamma}_n(m),\ n = 1, 2, \ldots\}$ converges in probability to γ.

It is clear that, having only condition (6) without any additional information about the rate of convergence of $\ln(V(tx)) - \ln(V(t))$ to $\gamma \ln(x)$, it is difficult to get good properties, such as the asymptotic normality, of any estimator of the parameter γ (or α). Therefore, from now we assume there exists a measurable function $A(t)$ with constant sign near infinity, which is not identically zero, and $A(t) \to 0$ as $t \to \infty$ such that

$$\lim_{t\to\infty} \frac{\ln(V(tx)) - \ln(V(t)) - \gamma \ln(x)}{A(t)} = \frac{x^{\varrho} - 1}{\varrho} \tag{8}$$

for all $x > 0$. Here $\varrho < 0$ is the so-called second order parameter. It is known that (8) immediately implies $|A(t)| \in RV_\varrho$. For the Fréchet distribution $G_{1/\gamma}(cx)$, $c > 0$, and

indeed only for this distribution, the numerator of the fraction on the left hand side of (8) is zero.

We introduce the constants

$$\begin{aligned}\lambda(\varrho) &= \frac{2^{\varrho}-1}{\varrho\ln(2)}\Gamma(1-\varrho),\\ d^2 &= \frac{1}{\ln^2(2)}\left(\frac{\pi^2}{3}-7\ln^2(2)+4\ln(2)\ln(3)-\frac{\ln^2(3)}{2}-\frac{3\mathrm{Li}_2(3/4)}{2}-\mathrm{Li}_2(1/3)\right.\end{aligned}$$

where $\mathrm{Li}_2(x)$ denotes the dilogarithm function, which is defined by the integral

$$\mathrm{Li}_2(x) = -\int_0^x u^{-1}\ln(1-u)\mathrm{d}u, \quad x \le 1.$$

The numerical value of the last constant is $d^2 \approx 1.11417$. Our main result is the following Theorem.

Theorem 2.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables with a distribution function F satisfying $F(1) = 0$ and (8). Suppose that a sequence $m \equiv m(n) \to \infty$ satisfies

$$\frac{nA^2(m)}{m} \quad \to \quad \nu^2 \in [0,\infty), \quad n \to \infty. \tag{9}$$

then

$$\sqrt{\frac{n}{m}}\left(\hat{\gamma}_n(m)-\gamma\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\nu\lambda(\varrho), \gamma^2 d^2\right), \quad n \to \infty \tag{10}$$

and

$$\sqrt{\frac{n}{m}}\left(\hat{\alpha}_n(m)-\alpha\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(-\alpha^2\nu\lambda(\varrho), \alpha^2 d^2\right), \quad n \to \infty, \tag{11},$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution.

If $X_1, \ldots, X_n$ is a sample from the Fréchet distribution $G_a(cx)$, $x > 0$, where c is a positive constant, then $\hat{\gamma}_n(m)$ is the unbiased estimator of the parameter γ for any fixed $m = 1, 2, \ldots$ Moreover,

$$\sqrt{n}\left(\hat{\gamma}_n(m)-\gamma\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \gamma^2 d^2 m\right), \quad \sqrt{n}\left(\hat{\alpha}_n(m)-\alpha\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \alpha^2 d^2 m\right)$$

as $n \to \infty$. For the both estimators asymptotic variance is minimal under choice $m = 1$.

Consider a non-stationary max-MA(q) process

$$X_t = \xi_t \vee (C_{1,t}\xi_{t-1}) \vee \cdots \vee (C_{q,t}\xi_{t-q}), \quad t = 0, 1, \ldots, \tag{12}$$

where ξ_t, $t \in \mathbb{Z}$ are i.i.d. random variables with a distribution F, satisfying $F(1) = 0$ and (6). Time depending weights $C_{k,t} \geq 0$, $0 \leq k \leq q$ are piecewise constant: $C_{k,t} = \tilde{C}_{k,j}$, if t $\in [t_j, t_j+1)$, where $t_0 = 0$ and $t_j < t_j+1$, $j \geq 0$. For example, taking $q = 2$ we get

$$X_t = \xi_t \vee \left(\tilde{C}_{1,k}\xi_{t-1}\right) \vee \left(\tilde{C}_{2,k}\xi_{t-2}\right), \quad t \in [t_k, t_{k+1}).$$

Further we assume that weights of the non-stationary max-MA(q) process satisfies condition $\tilde{C} < \infty$, where

$$\tilde{C} := \sup_{j \geq 0} \max\left\{1, \tilde{C}_{1,j}, \ldots, \tilde{C}_{q,j}\right\}. \tag{13}$$

A sequence X_t, $t \geq 0$ in (12) is q-dependent, i.e., any finite-dimensional random vectors $(X_{i1}, \ldots, X_{iu})$ and $(X_{j1}, \ldots, X_{ju})$, such that $j_1 - i_u > q$, are independent. If $C_{k,t} \equiv C_k$, then X_t, $t \geq 0$ in (12) forms a stationary max-MA(q) process (5). The last our result is the following Theorem.

Theorem 2.3. Let $X_1, \ldots, X_n$ be a sample from (12). Suppose that a sequence $m \equiv m(n) \to \infty$ satisfies the condition

$$\frac{n}{m(n)\tau(n)} \to \infty, \quad n \to \infty, \tag{14}$$

where $\tau(n)$ denotes the cardinality of a set $\mathcal{T}(n) = \{t_1, t_2, \ldots\} \cap \{1, 2, \ldots, n\}$. Then $\hat{\gamma}_n(m)$ is a weakly consistent estimator for the parameter γ.

Actually, $\tau(n)$ is a number of regime switches of a time series (12) in a time segment $[1, n]$. Condition (14) states that the number of groups $[n/m]$ tends to infinity faster than the number of regime switches $\tau(n)$.

Proofs of Theorem 2.1-2.3 are provided in Vaičiulis (2014).

REFERENCES

Davydov, Y., and Paulauskas, V. (1999). "On the Estimation of the Parameters of Multivariate Stable Distributions". *Acta Appl. Math.*, 58, 107-124.

De Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New York.

Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I., and Pestana, D. (2008). "Statistics of Extremes for iid Data and Breakthroughs in the Estimation of the Extreme Value Index: Laurens de Haan Leading Contributions". *Extremes*, 11, 3-34.

Paulauskas, V. (2003). "A New Estimator for Tail Index". *Acta Appl. Math.*, 79, 55-67.

Paulauskas, V., and Vaičiulis, M. (2011). "Several Modifications of DPR Estimator of the Tail Index". *Lith. Math. J.*, 51, 36-50. 2011.

Paulauskas, V., and Vaičiulis, M. (2012). "Estimation of the Tail Index in the Max-Aggregation Scheme". *Lith. Math. J.*, 52, 297-315. 2011.

Qi, Y. (2010). "On the Tail Index of a Heavy Tailed Distribution". *Ann. Inst. Statist. Math.*, 62, 277-289.

Vaičiulis, M. (2012). "Asymptotic Properties of Generalized DPR Statistic". *Lith. Math. J.*, 52, 95-110.

Vaičiulis, M. (2014). "Local Maximum Based Tail Index Estimator". To appear in *Lith. Math. J.*

97. APPLICATION OF STOCHASTIC MODELING TO ONLINE PRODUCT RETURNS RISK ANALYSIS

J.L. Vilar and A. Heras
Universidad Complutense de Madrid, Departamento de Economía Financiera y Actuarial, Spain,
jlvilarz@ccee.ucm.es
aheras@ccee.ucm.es

E. Vilar
Returnly Technologies Inc., San Francisco, USA,
eduardo@return.ly

ABSTRACT

One of the main customer's concerns with respect to on line shopping is the possibility of getting an unsatisfactory experience as a result of the absence of product evaluation and assessment. A straight forward growth sales strategy for on line stores is therefore preventing this to happen. Thus on line stores should view this phenomenon as a risk incurred by customers that might be transferred into the set of common risks supported by the on line retailer. This risk transfer consists firstly in adopting a product returns policy insuring the shoppers from any unsatisfactory experience and secondly efficiently managing it. In this communication we introduce some mathematical methodologies to assess and manage on line product returns risk. These consist basically in the application of stochastic modelling and risk analysis methods usually applied in the fields of Insurance and Finance (see for instance Tse, 2009; Dowd, 2006).

1. INTRODUCTION

The advent of Internet has allowed the quick expansion and adoption of online commerce. A variety of products and services are now sold, bought or exchanged through electronic platforms. Users are now able to buy goods and products that range from apparel, electronic devices, appliances, books and even perishable goods such as produce.

The online commerce process has inherent limitations such as the inability of consumers to touch and feel products, or their inability to accurately assess the size and quality of the product. Further it is complex to render a product, particularly aesthetic purchases such as apparel with 100% accuracy on an online commerce platform, an important aspect of product evaluation and assessment. Often when a user is unsatisfied with a received product, he will return the product to the merchant for exchange or reimbursement.

Product return is the return of goods or merchandise for a particular transaction, to a merchant, such as an individual or business, initiated by the shopper. Specifically it is the shipment and delivery of a product or merchandise from the shopper's address or shipping facility to a merchant's warehouse or product return location. In general, product returns can occur when, among other reasons, the product delivered to shoppers does not correspond to what they were expecting. For example, a product return may occur if the item is damaged or defective, if the order was accidental, if the product did not correspond to what the shopper had purchased (e.g. different size, color, or form), if the product is no longer wanted or if the order was not authorized by the shopper. The shopper may return the product for various reasons including others not listed above.

As logistic carriers and online commerce companies enable expedited deliveries, the number and frequency of product returns increases significantly. This increase in the number of product return is costly, unpredictable and adds complexity to the management of online commerce processes for the merchants. For example when a product return is requested by a shopper, it will involve the cost of shipping the product back to the merchant returns, a refund of money to the shopper and additional costs for re-packaging, restocking and reselling the returned item. In most

cases these additional costs were not factored into the price originally charged to the shopper, and consequently directly erodes the merchant's profit margins. It is at this point where the skill of financially managing these random costs appears to be critical for an on line retailer.

Thus product returns are unpredictable and can be shopper induced or merchant-induced. For example, the merchant may be responsible for damaged or defective items delivered, miss-picked or miss-packaged items etc. In addition, the fact that product returns can happen weeks or even months after the product was originally delivered to the shopper increases product return volatility, therefore increasing risk and making it hard for most retailers to budget for it. In many commerce verticals like apparel, getting items returned weeks after delivery can mean them being off-season or a huge drain on their underlying value. Thus return cost can be hard to quantify and will depend on sectors an considered costs. It is therefore desirable to be able to represent this random cost in a practical mathematical form able to allow the subsequent risk management. This is the part corresponding to risk assessment thanks to the application of stochastic modelling methodology.

Product return can also be costly and even considered a deterrent to purchase for the shopper. A product return policy can be a non-negligible incentive or deterrent for the final purchase decision (Mollenkopf, Rabinovich, Laseter and Boyer, 2007; Mollenkopf, Frankel and Russo, 2011; Ramanathan, 2011). In fact, the shopper often considers various criteria before arriving at a decision to purchase a product or not. For example, the shopper may consider delivery times, product return policies and shipping charges. Good product return policies may add to the confidence and comfort a shopper has with a merchant and will increase his loyalty and commitment to the merchant (Che, 1996; Griffis, Rao, Goldsby and Niranjan, 2012). Therefore the product returns phenomenon is a critical issue for on line retailers.

2. ONLINE PRODUCT RETURNS RISK ASSESSMENT AND MANAGEMENT

Product return risk consists in random a cash flow of future losses. To face them we will assume that the on line store may create a financial reserve nourished by a random cash flow consisting of some percentage (loading) of each sale income. We will

call this the income cash flow or shortly the premiums (by analogy with the insurance world). Return losses cash flow will be subtracted from this reserve. The difference in each period between these two cash flows will be called the surplus at the end of that period. Return risk assessment and management will consist in the probabilistic modeling and risk analysis of this surplus.

In this communication we proceed to the mathematical analysis of product returns risk in two steps.

The first step consists in building stochastic models able to supply the surplus cumulative distribution function (cdf). We first consider one period models where a period is typically equal to a month. In this case we obtain closed form formulae for the surplus cumulative distribution function. Next we consider the far more realistic multi period case, where we typically try to assess the risk over a one year horizon. This is a more complex case since the need of considering real world characteristics as trend and seasonality of either the transaction number and the return rate make it difficult to obtain a closed form formula for the surplus cdfs through many periods. Thus the need to consider a simulation model. As a result of this simulation methodology, in the multi period model we will still be able to produce samples of all the important variables (transactions, premiums, return losses, surplus, at any period) by means of a dynamic model simulation linking all variables from one period to the next.

In the second step we want to achieve risk measurement, i.e., to quantify return losses at some probability level. Two questions will be answered here. Firstly what percentage of each sale (i.e. the loading) is necessary to finance the return losses? By analogy to the insurance world, this is the pricing question. Secondly If the answer to the previous question was not affordable from a market competition point of view, we could still calculate our maximum (due to market competition conditions) feasible loading then trying to ask the following question: in addition to those insufficient premiums, what would be the initial amount of money needed to cover future random product return losses with a high probability? This is the economic capital question, again by analogy to the insurance and finance worlds. This two questions will be answered by means of the analysis of either the closed form surplus cdf (one period case) or the samples obtained by means of simulation (multi

period model). Depending on the case we will be able to calculate important figures to achieve risk analysis like loss ratios (Return Losses/premiums), survival probabilities, and also some risk measures such as Value at Risk and the conditional value at risk.

At the end of the second step the decision maker should be aware of how high the loading for financing product returns losses should be and how much capital should it be necessary to immobilize in case the necessary loading would not be fulfilled. This are the two important question to be asked in this communication for an efficient online product return risk management.

3. DATA SET

Fifty monthly transaction and return numbers were recorded from an on line store working in the USA apparel market beginning 2009-11 ending 2013-12. We apply to this real world example all the aforementioned models and ask to the pricing and economic capital questions basic questions.

4. CONCLUSION

We derive in a real world case the basic figures to manage on line product returns risk: the loadings and the economic capital necessary to face risk at a given probability level. This are derived by means of the application of some risk measures to the surplus cdfs, may this be obtained in closed form (one period case) or through a sample (multi period case) as numerous as our hardware might allow us.

REFERENCES

Che, Y. (1996). "Customer Return Policies for Experience Goods". *The Journal of Industrial Economics*, XLIV (1), 17-24.

Dowd, K. (2006). *Measuring Market Risk*. Wiley Finance.

Griffis, S.E., Rao, S., Goldsby, T.J., and Niranjan, T.T. (2012). "The Customer Consequences of Returns in Online Retailing: An Empirical Analysis". *Journal of Operations Management*, 30, 282-294.

Mollenkopf, D.A., Rabinovich, E., Laseter, T.M., and Boyer, K.K. (2007). "Manaing Internet Product Returns: A Focus on Effective Service Operations". D.S. Institute (ed.), *Decision Sciences*, 38 (2), 215-250.

Mollenkopf, D., Frankel, R., & Russo, I. (2011). "Creating Value through Returns Management: Exploring the Marketing-Operations Interface". *Journal of Operations Management* (29), 391-403.

Ramanathan, R. (2011). "An Empirical Analysis on the Influence of Risk on Relationships between Handling of Product Returns and Customer Loyalty in E-commerce". *International Journal of Production Economics* (130), 255-261.

Tse, Y. (2009). *Non Life Actuarial Models. Theory, Methods and Evaluation*. New York: Cambridge University Press.

98. DECOMPOSITION OF HAIR MINERAL VARIATIONS INTO INTRA-INDIVIDUAL, INTER-INDIVIDUAL AND PHYSICAL VARIATIONS

T. Yamada
Osaka University,School of Medicine, Japan,
t-yamada@stat.med.osaka-u.ac.jp

K. Kataoka
Mie University,School of Medicine, Japan,
kkoshi32@gmail.com

T. Saunders
Nagasaki University, School of Medicine, Japan,
styer2000@hotmail.com

K. Sera
Iwate Medical University, Cyclotron Research Center, Japan,
ksera@iwate-med.ac.jp

T. Takatsuji
Nagasaki University, School of Science and Engineering, Japan,
takatsuj@nagasaki- u.ac.jp

T. Nakamura
Chuo University,School of Science and Engineering, Japan,
naka@nagasaki-u.ac.jp

Y. Nose
Kumamoto Health Science University,School of Health Sciences, Japan,
noseyosi@oct- net.ne.jp

ABSTRACT

Since hair strands provide a record of past and current levels of mineral content in the body, they have been analyzed to study associations with onset of disease as well as with health status. Despite the wide body of scientific literature devoted to the study, hair mineral content is not regarded as a reliable measure in medicine, because of inconsistent and variable results. We conducted a cohort study of 834-mother-infant pairs to determine the association between hair minerals at one month and the onset of atopic dermatitis at ten-months after birth. In executing the study, we were faced with large intra-individual variations in most of the minerals. Since those large variations could likely degrade the associations between minerals and diseases, we hypothesized that ignoring the intra-individual variations might cause the unreliable and inconsistent results described in the literature. In this paper we describe the intra-individual variations of *Sr* and *Hg* among 209 6-year old children resampled 6 years after the original cohort, with special consideration concerning the measurement errors due to PIXE experimental conditions that were not well understood in either our previous reports or other literature.

1. INTRODUCTION

In 2005, we established a cohort sample of 834 infant-mother pairs in Fukuoka, Japan, and collected hair samples from them at the national one-month health checkup to measure hair mineral content using proton induced X-ray emission (PIXE) method (Sera *et al*., 2002). Associations between the hair mineral measurements at the one-month checkup and the onset of atopic dermatitis (AD) as diagnosed by pediatricians at the ten-month checkup were explored to obtain a model for detecting infants at high risk of developing AD. Of the 32 minerals measured, Selenium (*Se*) and Strontium (*Sr*) showed statistically significant associations with the onset of AD. These mineral amounts together with individual AD family history were incorporated into a logistic model to predict the risk of infant AD (Yamada *et al*., 2013). In executing the study, we were faced with large intra-individual variations among individual strands and measurement locations along the strands that could make clinical use of hair minerals problematic (Saunders *et al*., 2011).

In 2012, we resampled 209 then 6-year-old children from the cohort sample (n=834) to statistically analyze the variations in hair minerals in order to better promote their medical use. The objective of this paper is to decompose the large variations in *Sr* and *Hg* into those due to experimental error, and intra- and inter-individual variances.

2. METHODS

Each of the 209 children's hair strand samples was divided into two specimens for PIXE analysis to obtain two independent measurements.

Each mineral measurement $\tilde{X}_{ij}$ $(j=1,2;\ i=1,\ldots,n)$ is associated with standard error $\tilde{v}_{ij}$ due to physical variations, or simply referred to as physical error, in PIXE analysis. Since we found the square-root transformation useful (Yamada *et al*., 2015), $\tilde{X}_{ij}$ is square-root transformed. Let $X_{ij}=\tilde{X}_{ij}^{1/2}$ and the standard error for X_{ij} obtained from the delta method be denoted by v_{ij} .

Unbiased estimates for the variance due to the experimental error, the location, the inter- and intra-individual are presented in the Appendix.

Here we concern ourselves with *Sr* and *Hg*, since *Sr* is associated with the onset of the AD in our previous study (Yamada *et al*., 2013) and *Hg* has received a large amount of attention in scientific literature (Kempson *et al*., 2011). In the rest of the paper, "measurements" always indicate X_{ij} rather than $\tilde{X}_{ij}$ unless otherwise indicated.

3. RESULTS

3.1. *Sr*

Figure 1(a) and (b) show histograms of an average and a difference (D) of the two measurements of Sr, respectively, for 166 out of 209 subjects. Where, 43 pairs of measurements $(X_{i1}\ X_{i2})$ such that the estimated variance $\mathrm{V}(D_i)=(v_{i1}^{\ 2}+v_{i2}^{\ 2})$ of $D_i=X_{i1}-X_{i2}$ is larger than 0.8 are excluded since they are regarded as outliers in Figure 1(c),

the histogram of $V(D_i)$. Hair mineral content is subject to food and drug intake, disease, and other factors (Kempson *et al.*, 2011), so it stands to reason that outliers should be excluded in determining distributions of healthy subjects. This exclusion criterion is adopted in this study for the sake of simplicity and practicality. Figure 1(a) approximates a Weibull distribution with a test of fitness not significant (p=0.085). Figure 1(b) shows approximate normality with a test of fitness, again not significant (p=0.684). Figure 1 (d) shows a scatter plot for X_{i1} vs X_{i2} . The variance of crude inter-individual, intra individual σ_J^2, experimental error σ_e^2 , location error σ_L^2 and inter-individual variance σ_I^2 are 0.638, 0.152, 0.127, 0.025 and 0.486, respectively.

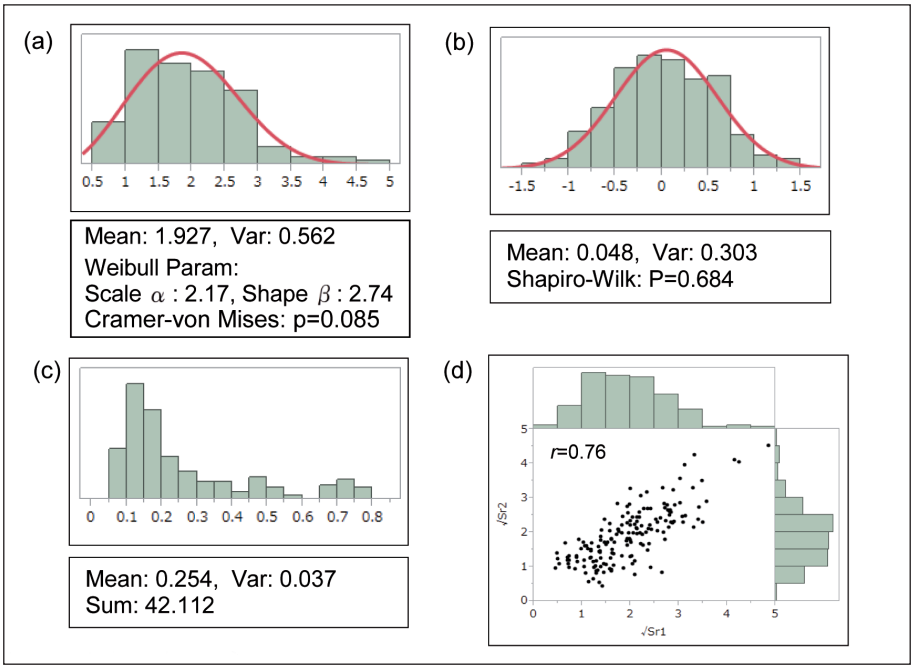


Figure 1. Results of *Sr*. (a) Histogram of Mean, (b) Histogram of Difference (D), (c) Histogram of Experimental Error Variance of D, (d) Scatter plot for $\sqrt{Sr_1}$ vs $\sqrt{Sr_2}$.

3.2. *Hg*

Figure 2 (a) ~ (d) show the results with *Hg* corresponding to Figure 1(a) ~ (d). Both Figure 2 (a) and (b) demonstrate approximate normality with the Shapiro-Wilk Test

not significant. Figure 2 (c) indicates the magnitude of the experimental errors about ten times larger than those for *Sr*, resulting in failure to calculate the variance due to locations. Since $\sigma_e^2 > \sigma_J^2$, the location error variance is not obtained. To obtain $\sigma_e^2 > \sigma_J^2$, we tried to exclude sufficient number of outliers based on $V(D_i)$ as in Sr, however this was not possible. This indicates the major cause of the intra-individual variation in *Hg* is due to PIXE experiment, as most *Hg* measurements are under the detection limit. This is understandable, due to *Hg* is toxicity.

The Crude inter-individual variance, intra-individual variance σ_J^2 and experimental error variance σ_e^2 are 0.304, 0.227 and 0.314, respectively.

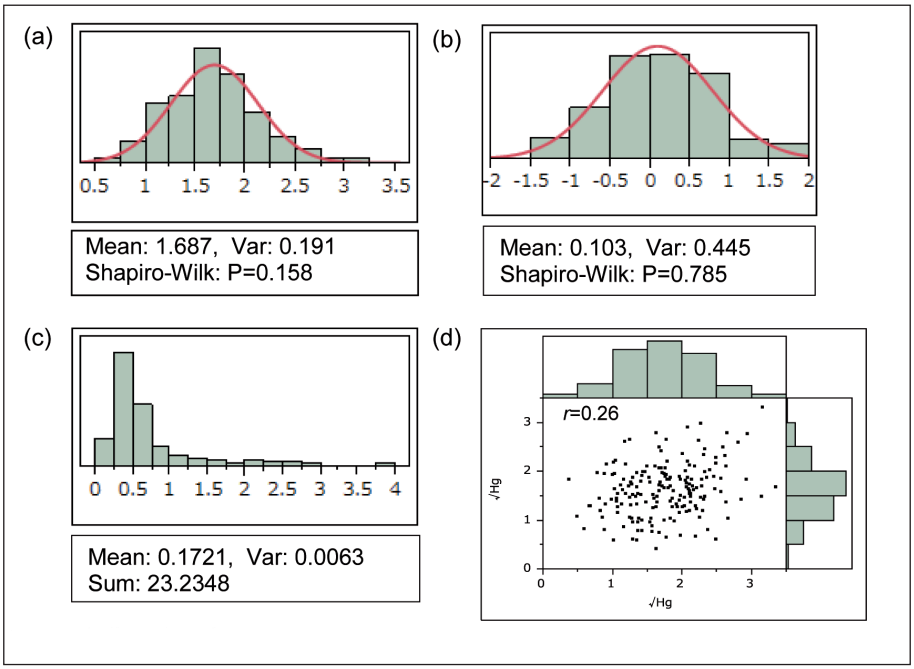


Figure 2. Results of *Hg*. (a) Histogram of Mean, (b) Histogram of Difference (D), (c) Histogram of Experimental Error Variance of D, (d) Scatter plot for $\sqrt{Hg_1}$ vs $\sqrt{Hg_2}$.

4. DISCUSSION

A systematic review of the literature presenting reference values of minerals in human hair emphasized that the reference values vary and that it is necessary to clarify standard procedures to further validate hair mineral analysis. There exist many exogenous or endogenous variables, besides age and gender that affect concentrations such as external environment and geography (Kempson *et al.*, 2011). As a result, outliers are inevitable for some minerals. These outliers should be excluded when establishing normal ranges. We propose a criterion for excluding those outliers to better understand the causes of the variations. Researchers should take into account the intra-individual variations and experimental errors as well as inter-individual variations particularly for medically important minerals such as *Sr* or *Hg*, since those minerals show extremely large variations.

We obtained two independent measurements from each subject to decompose the hair mineral variations into inter-individual and intra- individual variations. The intra-individual variation was further decomposed into variations due to location and those due to experimental error. The proposed methods work well with *Sr*, but not for *Hg*. The intra-individual variations for *Sr* and *Hg* are both normal and therefore regarded as random error. However, the experimental error for *Hg* is too large to detect the location error variance. To obtain more reliable estimates of mineral amounts with smaller experimental errors, it would be interesting to apply a maximum likelihood method to X-ray spectral data (Sera *et al.*, 2002) to remove background X-ray counts, however this is left for future work.

APPENDIX: MODELING OF INTRA- AND INTER- INDIVIDUAL VARIANCE

We have two mineral measurements $\tilde{X}_{ij}$ $(j=1,2)$ obtained from a specimen of hair strands i *(*$i = 1, \ldots, n$*)* that are measured at different locations. It is assumed that those measurements are mutually independent given each specimen. Each measurement $\tilde{X}_{ij}$ is associated with a standard error $\tilde{v}_{ij}$ due to the physical conditions in PIXE analysis. We denote by X_{ij} and v_{ij} the square root transformation of $\tilde{X}_{ij}^{1/2}$ and the

standard error of X_{ij} obtained from the delta method, or $v_{ij}^2 = \tilde{v}_{ij}^2 / 4\tilde{X}_{ij}$, respectively. We assume the following simple additive model:

$$X_{ij} = \mu_i + \delta_{ij} \text{ with } \delta_{ij} = \varepsilon_{ij} + e_{ij} .$$

μ_i true or exact value to denote the average of the mineral amount of all locations among the hair strands of i^{th} subject and $V(\mu_i) = \sigma_I^2$ is termed adjusted inter-individual variance.

δ_{ij} deviation from μ_i . $E(\delta_{ij}) = 0$ and $V(\delta_{ij}) = \sigma_J^2$ is intra-individual variance. $V(X_{ij}) = \sigma_I^2 + \sigma_J^2$ is termed crude inter-individual variance.

ε_{ij} deviation from μ_i due to location where hair strands are sampled. $\varepsilon_{ij} \sim N(0, \sigma_L^2)$ and σ_L^2 is *variance due to location.*

e_{ij} measurement error due to physical conditions of PIXE analysis. $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_e^2$ is variance due to experimental condition.

v_{ij}^2 unbiased estimate of the variance $V(e_{ij})$ obtained from PIXE analysis.

It is assumed ε_{ij} and e_{ij} are mutually independent, and δ_{ij}, ε_{ij} and e_{ij} are uncorrelated with μ's . Then it holds $V(\delta_{ij}) = \sigma_J^2 = \sigma_L^2 + \sigma_e^2$.

We assume $D_i = X_{i1} - X_{i2}$ is approximately normally distributed. It follows from the law of large numbers that

$$(2n)^{-1} \sum_{i=1}^{n} D_i^2 = (2n)^{-1} \sum_{i=1}^{n} (\delta_{i1} - \delta_{i2})^2 \cong \sigma_J^2 \quad (1)$$

where $\cong$ denotes asymptotic equivalence.

Since $E[(\varepsilon_{i1} - \varepsilon_{i2})^2] = \sigma_L^2$, $E[(e_{i1} - e_{i2})^2] = v_{i1}^2 + v_{i2}^2$ and $E[(\varepsilon_{i1} - \varepsilon_{i2})(e_{i1} - e_{i2})] = 0$, we have

$$(2n)^{-1} \sum_{i=1}^{n} D_i^2 \cong \sigma_L^2 + (2n)^{-1} \sum_{i=1}^{n} (v_{i1}^2 + v_{i2}^2) , \text{ or}$$

$$\sigma_L^2 \cong (2n)^{-1} \sum_{i=1}^{n} D_i^2 - (2n)^{-1} \sum_{i=1}^{n} (v_{i1}^2 + v_{i2}^2) \quad (2)$$

The second term in the right hand side is asymptotically equal to σ_e^2.

Put $Y_i = \frac{X_{i1} + X_{i2}}{2}$, $\bar{Y} = \frac{1}{n}\sum_i Y_i$, $\bar{\delta}_i = \frac{\delta_{i1} + \delta_{i2}}{2}$ and $\bar{\bar{\delta}} = \frac{1}{n}\sum_i \bar{\delta}_i$ then

$Y_i - \bar{Y} = (\mu_i - \bar{\mu}) + \left(\bar{\delta}_i - \bar{\bar{\delta}}\right)$. It holds that

$$(n-1)^{-1}\sum_i \left(Y_i - \bar{Y}\right)^2 = (n-1)^{-1}\sum_i (\mu_i - \bar{\mu})^2 + (n-1)^{-1}\sum_i \left(\bar{\delta}_i - \bar{\bar{\delta}}\right)^2$$

The sums of squares in the right hand side asymptotically equal σ_I^2 and $\sigma_J^2/2$, respectively, and thus

$$(n-1)^{-1}\sum_i \left(Y_i - \bar{Y}\right)^2 \cong \sigma_I^2 + \left(\sigma_J^2/2\right), \text{ or}$$
$$\sigma_I^2 \cong (n-1)^{-1}\sum_i \left(Y_i - \bar{Y}\right)^2 - \left(\sigma_J^2/2\right) \quad (3)$$

Crude inter-individual variance is $\sigma_I^2 + \sigma_J^2 \cong (n-1)^{-1}\sum_i \left(Y_i - \bar{Y}\right)^2 + \left(\sigma_J^2/2\right)$.

ACKNOWLEDGEMENTS

Special thanks to the Nishina Memorial Cyclotron Center staff and all the mothers and infants who participated.

REFERENCES

Kempson, I.M., and Lombi, E. (2011). "Hair Analysis as a Biomonitor for Toxicology, Disease and Health Status". *Chemical Society Reviews*, 40, 3915-3940.

Saunders, T., Kuroda, S., Makie, T., Sera, K., Yamada, T., Nakamura, T., and Nose, Y. (2011). "Effect of Biological Variability Incidental to PIXE-Hair Minerals on Risk Analysis of Atopic Dermatitis". *Biometrie und Medizinische lnformatik-Greifswalder Seminarberichte: Risk Analysis*, 18, 127-137.

Sera, K., Futatsugawa, S., and Murao, S. (2002). "Quantitative Analysis of Untreated Hair Samples for Monitoring Human Exposure to Heavy Metals". *Nuclear Instrum. and Meth. in Physics Research*, B189, 174-179.

Yamada, T., Saunders, T., Kuroda, S., Sera, K., Nakamura, T., Takatsuji, T., Hara, T., and Nose, Y. (2013). "Cohort Study for Prevention of Atopic Dermatitis using Hair Mineral Contents". *J. Trace Elem. Med. Biol.*, 27 (2), 126-131.

Yamada, T., Saunders, T., Nakamura, T., Sera, K., and Nose, Y. (2015). *On Intra-Individual Variations in Hair Minerals in relation to Epidemiological Risk Assessment of Atopic Dermatitis*. Springer, to appear.

99. ON A SHAPE-INVARIANT HAZARD REGRESSION MODEL

C. Zheng
University of Wisconsin, Milwaukee, Wisconsin, USA
zhengc@uwm.edu

Y.Q. Chen
Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
yqchen@fhcrc.org

ABSTRACT

In survival analysis, the Cox model is widely used for most clinical trial data. Alternative models include the additive hazard model, the accelerated failure time (AFT) model and a more general transformation model. All these models assume that the effects for all covariates are on the same scale. However, it is possible that for different covariates, the effects are on different scales. In this paper, we propose a shape-invariant hazard regression model to allow all covariate effects on different scales. We propose moment-based inference procedures for the regression parameters. We also discuss the risk prediction and goodness of fit test for our proposed model. Numerical simulation studies show that the proposed estimator has a good performance.

1. INTRODUCTION

In time-to-event analysis, the Cox model (Cox, 1972) is widely used in risk assessment to identify risk factors via regression. Alternative models exist. For example, one alternative is the additive risk model (Lin and Ying, 1994), which assumes that the covariates' effect on hazard functions is additive, rather than multiplicative. Another important alternative model is the accelerated failure time model (AFT;

Zeng and Lin, 2007), whose coefficient has better scientific explanation and is commonly used in the field of engineering or physics, i.e., reliability assessment. Additional alternatives include a broader class of transformation models, such as the proportional odds model. Nevertheless, all these models assume that all the covariates have the same type of effect on a unified transformed scale. In reality, it is possible that different risk factors have different kinds of effects, while modeling all variables' effects in the same scale might not be optimal.

Several models have been proposed to model different kinds of effects. Lin and Ying (1995) proposed an additive-multiplicative hazard model that allows two kinds of effects. Chen and Jewell (2001) proposed an accelerated hazard model which is a flexible model allowing both effect in multiplicative hazard scale and time scale change. In this paper, we propose a broader class of model that allows all three kinds of effects. To make the interpretation of each parameter clear, we focus our method on the cases where covariates are different for different effects in the main part of the paper and then discuss how our methods can be extended to the case where same covariate have different kind of effects. Besides estimating the effect of certain covariate, as a complex model, our model is useful to make prediction since our model is more closer to a nonparametric one so that it has few restrictions between the survival curves for different groups defined by covariates. We organize the paper as follows. In §2, we propose a shape invariant hazard model. In §3, we summarize the inference method with asymptotic results. In §4, we present some numerical results to evaluate the performance of our estimator. In §5, we give more discussion of the model.

2. THE SHAPE-INVARIANT HAZARD REGRESSION MODEL

Assume $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$ are time independent risk factors, β's are regression parameters, and λ_0 is an unknown baseline hazard function. Then the "shape-invariant model" we considered is in the following general form:

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_2} + \boldsymbol{\beta}_3 \boldsymbol{Z}_3. \tag{1}$$

More generally, when $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ are potentially time-varying, we have the model in the following form:

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}} + \boldsymbol{\beta}_3 \boldsymbol{Z}_3(t). \tag{2}$$

We use this kind of model since it allows the survival functions to have cross-over, which is the case for many real data problem. Here $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2(t)$, $\boldsymbol{Z}_3(t)$ are assumed to be different. So for this model, we have an accelerated time effect from $\boldsymbol{Z}_1$, then a multiplicative hazard effect from $\boldsymbol{Z}_2(t)$ and finally an additive effect from $\boldsymbol{Z}_3(t)$. An important feature for this model is that for any subgroup defined by $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, the model is an additive risk model with respect to $\boldsymbol{Z}_3$. But the model is not a Cox model for some subgroups defined by $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_3$. It is only a Cox model for the subgroup with $\boldsymbol{Z}_3 = 0$. Also, the model is an AFT model only in the subgroup with $\boldsymbol{Z}_2 = 0$ and $\boldsymbol{Z}_3 = 0$. The reason for this phenomenon is due to the order of including three kinds of effect. We can also choose different orders for the effects and obtain the following models:

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}} + \boldsymbol{\beta}_3 \boldsymbol{Z}_3(t) e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}}, \tag{3}$$

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}} + \boldsymbol{\beta}_3 \boldsymbol{Z}_3(t) e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1}, \tag{4}$$

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}} + \boldsymbol{\beta}_3 \boldsymbol{Z}_3(t) e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1 + \boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}}, \tag{5}$$

In most case, we would like to have a treatment effect on the Cox part. Since the model can be viewed as a stratified proportional hazard model for model (2.3) and (2.5), we can directly fit the Cox model without using our proposed complex model. So the method for estimating parameters in model (2.2) and (2.4) may be more meaningful in practice. Since the estimation procedures are similar for these models, we focus on model (2.2) in this paper.

To make β in model (2.2) identifiable, we need that for some t, $[\lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1}) + te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1} \dot{\lambda}_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})]\boldsymbol{Z}_1 e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1 + \boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}}$, $\lambda_0(te^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1})\boldsymbol{Z}_2(t) e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1 + \boldsymbol{\beta}_2 \boldsymbol{Z}_{2(t)}}$, and $\boldsymbol{Z}_3(t)$ are non-collinear at true parameter. When the joint distribution of $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, $\boldsymbol{Z}_3$ are nondegenerate, this assumption often holds.

The model has the cumulative hazard function in the following form:

$$\Lambda(t \mid \boldsymbol{Z}) = \int_0^t [e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2(u)}} d\Lambda_0(u e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1}) + \boldsymbol{\beta}_3 \boldsymbol{Z}_3(u) du]. \quad (6)$$

When $\boldsymbol{Z}_2$ is not time dependent, we have

$$\Lambda(t \mid \boldsymbol{Z}) = e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_2} \Lambda_0(t e^{\boldsymbol{\beta}_1 \boldsymbol{Z}_1}) + \boldsymbol{\beta}_3 \int_0^t \boldsymbol{Z}_3(u) du. \quad (6)$$

3. MOMENT-BASED INFERENCE PROCEDURES

In this section, we give the inference procedure for our proposed hazard shape invariant model. We first give the estimator, followed by the asymptotic result. After that, we discuss the prediction and test for goodness fit of the model.

We denote our data collected as $\underline{X} = \{Y_i(t), N_i(t), \boldsymbol{Z}_i(t), t \in [0, \infty), i = 1,2, \ldots, n\}$. We define the filtration F_t as

$$\sigma\{Y_i(s \exp(-\boldsymbol{\beta}_{10} \boldsymbol{Z}_{1i})), N_i(s \exp(-\boldsymbol{\beta}_{10} \boldsymbol{Z}_{1i})), \boldsymbol{Z}_{1i}, \boldsymbol{Z}_{2i}(s), \boldsymbol{Z}_{3i}(s), s \le t, i \in \mathcal{N}\}.$$

The true parameters are $\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30})$. We thus derive the moment based estimator by noticing $E[dN_i(te^{-\boldsymbol{\beta}_{10} \boldsymbol{Z}_{i1}}) | \mathcal{F}_{t-}; \boldsymbol{Z}_i, \boldsymbol{\beta}_0]$ equals

$$Y_i(te^{-\boldsymbol{\beta}_{10} \boldsymbol{Z}_{1i}})[\lambda_0(t) e^{\boldsymbol{\beta}_{20} \boldsymbol{Z}_{2i}(t)} + \boldsymbol{\beta}_{30} \boldsymbol{Z}_{3i}(t) e^{-\boldsymbol{\beta}_{10} \boldsymbol{Z}_{1i}}] dt.$$

So we can obtain an estimator for baseline hazard function given regression parameters as

$$\widehat{\Lambda}_0(t; \boldsymbol{\beta}) = \int_0^t \frac{\sum_i [dN_i(se^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}}) - Y_i(se^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}}) \boldsymbol{\beta}_3 \boldsymbol{Z}_{3i}(s) e^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}} ds]}{\sum_i Y_i(se^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}}) e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2i}(s)}}.$$

Let $\overline{W}(t; \boldsymbol{\beta}) = \frac{\sum_j W_j(t) Y_j(te^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1j}}) e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2j}(t)}}{\sum_{1j} Y_j(te^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1j}}) e^{\boldsymbol{\beta}_2 \boldsymbol{Z}_{2j}(t)}}$, then we can choose different $W_i(t)$ and solve the following unbiased estimating equations

$$\sum_{i=1}^n \int \{[W_i(t) - \overline{W}(t; \boldsymbol{\beta})][dN_i(te^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}}) - Y_i(te^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}}) \boldsymbol{\beta}_3 \boldsymbol{Z}_{3i}(t) e^{-\boldsymbol{\beta}_1 \boldsymbol{Z}_{1i}} dt]\} \quad (8)$$

to obtain a consistent estimator. When $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ are different, it is natural to use them as the weight function.

For simplicity, we can assume the weight does not depend on unknown parameters and empirical distribution of the survival time for whole population, i.e. we assume $W_i(t) = f(Z_i, t)$.

Theorem 1: Under certain regularity conditions, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma).$$

This theorem can be established following similar approaches in Tsiatis (1990).

4. MODEL-BASED PREDICTION

For prediction purposes, similar to the Cox model (Lin *et al.*, 1994), the AFT model (Park and Wei, 2003) and the additive risk model (Shen and Cheng, 1999), pointwise confidence interval as well as the simultaneous confidence bands can be constructed. For pointwise confidence interval, as we have the central limit theory for $\widehat{\beta}$ and $\widehat{\Lambda}(t)$, the only thing we need to know is the joint distribution of these two quantities to obtain a standard error estimation $v(t) = \widehat{SE}$ of our estimated cumulative hazard function

$$\widehat{\Lambda}(t|\boldsymbol{Z}) = \int_0^t [e^{\widehat{\boldsymbol{\beta}}_2 \boldsymbol{Z}_{2(u)}} d\widehat{\Lambda}_0(u e^{\widehat{\boldsymbol{\beta}}_1 \boldsymbol{Z}_1}) + \widehat{\boldsymbol{\beta}}_3 \boldsymbol{Z}_3(u) du],$$

where the baseline hazard is estimated by $\widehat{\Lambda}_0(t) = \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$. When $\boldsymbol{Z}$ is time independent, this can be simplified by

$$\widehat{\Lambda}(t|\boldsymbol{Z}) = \widehat{\Lambda}_0(t e^{\widehat{\boldsymbol{\beta}}_1 \boldsymbol{Z}_1}) e^{\widehat{\boldsymbol{\beta}}_2 \boldsymbol{Z}_2} + \widehat{\boldsymbol{\beta}}_3 \boldsymbol{Z}_3 t.$$

The pointwise confidence interval can be calculated by $\widehat{\Lambda}(t|\boldsymbol{Z}) \pm z_{1-\alpha/2} v(t)$, where z is the critical value from standard normal distribution and α is significant level. We can obtain $\widehat{v}(t)$ from simple Bootstrap method. Similarly, we can have the following type of confidence bands $\widehat{\Lambda}(t|\boldsymbol{Z}) \pm c(\alpha) v(t)$,where $c(\alpha)$ is obtained from Bootstrap sample.

5. MODEL ADEQUACY ASSESSMENT

We consider the Kolmogorov-Smirnov Test and the Gill-Schumacher Test in this section.

We can compare the model based fitted cumulative hazard to the nonparametric fitted one. This method is not preferred when $\boldsymbol{Z}$ is time dependent.

When $\boldsymbol{Z}$ is time independent, we define the test statistics as

$$\begin{aligned} & D(t,\widehat{\boldsymbol{\beta}},\boldsymbol{Z}) = n^{-1/2}\int_0^t Q(u,\widehat{\boldsymbol{\beta}},\boldsymbol{Z})d[\widehat{\Lambda}_M(u|\boldsymbol{Z}) - \widehat{\Lambda}_{NP}(u|\boldsymbol{Z})] \\ = \;& n^{-1/2}\int_0^t Q(u,\widehat{\boldsymbol{\beta}},\boldsymbol{Z})d[\widehat{\Lambda}_0(ue^{\widehat{\boldsymbol{\beta}}_1\boldsymbol{Z}_1})e^{\widehat{\boldsymbol{\beta}}_2\boldsymbol{Z}_2} + \widehat{\boldsymbol{\beta}}_3\boldsymbol{Z}_3 u - \widehat{\Lambda}_{NP}(u|\boldsymbol{Z})], \end{aligned}$$

where $\widehat{\Lambda}_{NP}(u|\boldsymbol{Z})$ can be obtained either from a Kernel-Smoothed Kaplan-Meier estimator when $\boldsymbol{Z}$ is in lower dimension and is continuous with form

$$\widehat{\Lambda}_{NP}(u|\boldsymbol{z}) = -\log\{\Pi_{T_i\le u,\Delta_i=1}\left[1 - \frac{K_a\{\widehat{F}(\boldsymbol{z}) - \widehat{F}(\boldsymbol{Z}_i)\}}{\sum_{j=1}^n I(\widehat{\boldsymbol{Z}}_j > \widehat{\boldsymbol{Z}}_i)K_a\{\widehat{F}(\boldsymbol{z}) - \widehat{F}(\boldsymbol{Z}_i)\}}\right]\},$$

or from a subgroup Kaplan-Meier estimator when $\boldsymbol{Z}$ is discrete. We can use $D_1(\widehat{\boldsymbol{\beta}}) = sup_{t,\boldsymbol{Z}}|D(t,\widehat{\boldsymbol{\beta}},\boldsymbol{Z})|$ or $D_2(\widehat{\boldsymbol{\beta}}) = sup_t|\int D(t,\widehat{\boldsymbol{\beta}},\boldsymbol{Z})d\widehat{F}_{\boldsymbol{Z}}|$ where $\widehat{F}_{\boldsymbol{Z}}$ is the empirical distribution of $\boldsymbol{Z}$. We can use some well known weight for $Q(u,\widehat{\boldsymbol{\beta}},\boldsymbol{Z})$. And a very large test statistics suggest that the model is not adequate to fit the data.

For any weight $W_{1i}(t)$ and $W_{2i}(t)$, we can obtain estimators by solving the score equation $S_1(\beta) = 0$ and $S_2(\beta) = 0$ to obtain $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$. It is straight-forward to use the test statistics $\widehat{\boldsymbol{\beta}}^{(1)} - \widehat{\boldsymbol{\beta}}^{(2)}$, but the variance of this statistics is difficult to compute. We consider the score test. Since we have that under the null, $n^{-1/2}(S_1(\beta_0), S_2(\beta_0))$ follows joint normal distribution with covariance matrix Σ_{12}, which can be estimated using Bootstrap method. So we can use test statistics

$$T_{GS} = min_{\beta\in U(\widehat{\boldsymbol{\beta}}^{(1)})}\left[\begin{pmatrix} S_1(\boldsymbol{\beta}) \\ S_2(\boldsymbol{\beta}) \end{pmatrix}^T \widehat{\Sigma}_{12}^{-1} \begin{pmatrix} S_1(\boldsymbol{\beta}) \\ S_2(\boldsymbol{\beta}) \end{pmatrix}\right],$$

which follows χ_p^2 distribution asymptotically to assess model fitting.

6. NUMERICAL STUDIES

To assess the sample performance of our proposed estimators, we have performed simulations as shown below. We chose the baseline hazard as $1/(1+t)$. We considered $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$ where $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ independently follows bernoulli distribution with parameter 0.5. The censoring distribution is an exponential distribution to making about 30% censoring. We ran 100 simulations for each setting and summarized the bias, empirical SE, mean estimated SE. The simulation results are summarized in Table 1. From the table, we can see that our proposed estimator is unbiased. The estimated SE is generally comparable to the empirical one and the empirical SE increases when the censoring rate increases.

	Bias			Est SE			Emp SE		
True β	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0,0,0.1	0.01	0.02	-0.01	0.20	0.06	0.03	0.18	0.10	0.05
0,0,0.2	0.01	0.02	-0.01	0.26	0.06	0.04	0.20	0.11	0.06
0,-0.5,0.1	0.01	0.01	0.00	0.24	0.07	0.03	0.20	0.10	0.04
0,-0.5,0.2	0.01	0.01	0.00	0.22	0.07	0.03	0.21	0.11	0.05
0.5,0,0.1	0.01	0.02	-0.01	0.18	0.05	0.03	0.18	0.10	0.06
0.5,0,0.2	0.00	0.02	-0.01	0.20	0.06	0.05	0.19	0.10	0.07
0.5,-0.5,0.1	0.00	0.01	0.00	0.22	0.07	0.03	0.20	0.10	0.05
0.5,-0.5,0.2	0.00	0.01	-0.01	0.22	0.09	0.05	0.21	0.11	0.06

Table 1. Simulation results.

7. CONCLUSION

We propose a general hazard regression model to assess risk factors on different scales. Our proposed estimation procedure is reasonable, simple and easy to implement.

REFERENCES

Chen, Y.Q., and Jewell, N.P. (2001). "On a General Class of Semiparametric Hazards Regression Models". *Biometrika* 88, 687-702.

Cox, D.R. (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society*, Series B 34, 187-220.

Lin, D.Y., and Ying, Z. (1995). "Semiparametric Analysis of General Additive-Multiplicative Hazard Models for Counting Processes". *The Annals of Statistics*, 23, 1712-1734.

Lin, D.Y., and Ying, Z. (1994). "Semiparametric Analysis of the Additive Risk Model". *Biometrika* 81, 61-71.

Park, Y., and Wei, L.J. (2003). "Estimating Subject-Specific Survival Functions Under Accelerated Failure Time Model". *Biometrika*, 90, 717-723.

Shen, Y., and Cheng, S.C. (1999). "Confidence Bands for Cumulative Incidence Curves Under the Additive Risk Model". *Biometrics*, 55, 1093-1100.

Tsiatis, A.A. (1990). "Estimating Regression Parameters Using Linear Rank Tests for Censored Data". *Annals of Statistics* 18, 354-372.

Zeng, D., and Lin, D.Y. (2007). "Efficient Estimation for the Accelerated Failure Time Model". *Journal of the American Statistical Association* 102, 1387-1396.

LIST OF AUTHORS

Abad, P.: 23
Abella, J.: 579
Aceves Martins, M.: 765
Acharyya, M.: 407
Adachi, T.: 473
Afonso, P.: 431
Alaminos, E.: 33
Albrecher, H.: 41
Alcañiz, M.: 49
Alemany, R.: 629
Andrés-Sánchez, J. de 57
Aranda, F.: 65
Arató, M.: 71
Armero, C.: 711
Arnold, B.C.: 79
Arriaza-Gómez, A.: 87
Ayestaran, A.: 367
Ayuso, M.: 33, 95
Baena Mirabete, S.: 103
Barranco-Chamorro, I.: 111
Bartoszczuk, P.: 119
Belles-Sampera, J.: 127
Belzunce, F.: 87
Bergfjord, O.: 135
Bermúdez, L.: 143, 153
Bistoni, N.: 621
Blanco Izquierdo, V.: 605
Boira, E.: 161
Boj, E.: 169
Bolancé, C.: 571
Cabaña, A.: 179
Calderín-Ojeda, E.: 377
Calvet, L.: 187
Cardis, E.: 525
Carnicer, A.: 525
Caroni, C.: 195, 203
Carracedo, P.: 211
Casero-Alonso, V.: 219
Castañer, A.: 229
Castillo, J. del 579
Cazorla, F.J.: 579
Chen, Y.Q.: 827
Chulia, H.: 23, 237
Claramunt, M.M.: 229
Corral, Á.: 245, 703
Cortés, J.C.: 257
Costa, M.T.: 169
Daily-Amir, D.: 41
Daoudi, J.: 251
Debón, A.: 211, 257, 773
Deodatis, G.: 491
Di Bernardino, E.: 265, 273
Economou, P.: 195, 203
Fernández-Ponce, J.M.: 265

Ferrando, P.: 281
Ferreiro-Castilla, A.: 289
Filus, J.: 297, 303
Filus, L.: 303
Forte, A.: 711
Fortiana, J.: 281
Francisco, C.: 311
Franco, G.: 491
Freixas, J.: 319
Gabarro, J.: 329
García, C.: 65
García, V.: 377
Ghanmi, A.: 729
Ghosh, I.: 79
Giebel, S.: 339, 347
Gil-Lafuente, A.M.: 483
Giralt Batista, M.: 765
Gómez Mateu, M.: 357
Gómez Melis, G.: 357, 367, 711
Gómez-Déniz, E.: 377, 613
González-Vila Puchades, L.: 57
Grellier, J.: 525
Guillén, M.: 143, 237
Gutiérrez-Miravete, E.: 385
Haberman, S.: 211
Hampel, D.: 391
Heras, A.: 811
Hermann, P.: 339
Herrero, R.: 399
Hobert, S.: 515
Houston, J.: 407
Huber, C.: 415
Hwang, I.G.: 757
Hwang, M.-S.: 757
Issever Grochová, L.: 423, 647
Janová, J.: 391
Jiménez, V.: 431
Jordá, V.: 695
Jordanova, P.: 439
Juan, Á.A.: 187, 457, 587, 655, 773
Kachalov, R.: 721
Kadlec, J.: 391
Karlis, D.: 143, 153
Kataoka, K.: 817
Kitsos, C.: 449
Kizys, R.: 457
Lambert, M.: 637
Laniado, H.: 795
Leal, C.: 551
Leo, P.Y.: 465
Li, W.: 473
Lillo, R.E.: 251, 677, 787, 795
Linares, A.: 457
Linares-Mustarós, S.: 483
Llauradó Ribé, E.: 765
Lopeman, M.: 491
López-Fidalgo, J.: 219
Malefaki, S.: 195
Markovich, L.: 499
Markovich, N.: 507
Mármol, M.: 229
Martinek, L.: 71

Mengelkamp, A.: 515
Merigó-Lindahl, J.M.: 483
Monzó-Chafer, O.: 257
Moreno, M.: 637
Moriña, D.: 525, 765
Mulero, J.: 87
Nakamura, T.: 817
Nefedova, Y.: 439
Neves, M.: 533
Nose, Y.: 817
Novak, S.: 541
Nunes, S.: 533
Oliveira, A.: 551
Oliveira, T.: 311, 551
Oramas Dorta, D.: 561
Ornelas, A.: 571
Orriols, G.: 399
Padilla, M.: 579
Páez, Á.: 711
Pagès Bernaus, A.: 587
Palacios Rodríguez, F.: 265
Panagoulia, D.: 203
Paulauskas, V.: 597
Penalva, H.: 533
Pérez Sánchez, J.M.: 65, 605, 613
Pérez-Marín, A.M.: 95
Pernot, E.: 525
Perpiñán, H.: 711
Pi, F.: 399
Pierri, F.: 621
Piulachs, X.: 629, 711
Platania, F.: 637
Plecitá, K.: 647
Pons, M.: 319
Pons, M.A.: 161
Prieto, F.: 687, 695
Puig, P.: 103
Quintero-Araujo, C.: 655
Rainer, M.: 347
Ramalhinho, H.: 587
Ramírez-Cobo, P.: 677
Ramon, L.: 49
Remuzgo, L.: 663
Rochet, P.: 671
Rodríguez, J.: 677
Rodríguez-Griñolo, M.R.: 265
Rodríguez-Santos, F.: 257
Rué, M.: 711
Rullière, D.: 273
Sáez-Castillo, A.J.: 687
Santolino, M.: 49, 153
Sarabia, J.M.: 663, 687, 695
Sarrasí, F.J.: 161
Saunders, T.: 817
Sawik, B.: 457
Schefers, N.: 187
Schenk, J.-P.: 339
Schumann, M.: 515
Sera, K.: 817
Serna, M.: 329
Serra, I.: 179, 245, 605, 703
Serrat, C.: 711

Sleptsova, Y.: 721
Sokri, A.: 729
Solà Alberich, R.: 765
Sordo, M.A.: 737
Stanghellini, E.: 621
Stehlík, M.: 339, 439, 741
Střelec, L.: 749
Suárez-Llorens, A.: 87
Suh, J-H.: 757
Takatsuji, T.: 817
Tarro, L.: 765
Terradez, M.: 773
Ting Lee, M.-L.: 781
Torrado, N.: 787
Torres Díaz, R.A.: 795
Toulias, T.: 449
Trueba, C.: 663
Uribe, J.: 237
Vaičiulis, M.: 597, 803
Vázquez-Polo, F.: 377
Vilar, E.: 811
Vilar, J.L.: 811
Yamada, T.: 817
Zheng, Z.: 827

COLECCIÓN "CUADERNOS DE LA FUNDACIÓN"

Para cualquier información sobre nuestras publicaciones consulte:
www.fundacionmapfre.org

205. Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference. 2015

204. Determinantes do Prêmio de *Default* de Crédito de (Res)Seguradores. 2014

203. Generación de escenarios económicos para la medición de riesgos de mercado en Solvencia II a través de modelos de series temporales. 2014

202. La valoración de los inmuebles del patrimonio histórico y los riesgos sísmicos en el contrato de seguro: el caso de Lorca. 2014

201. Inteligencia computacional en la gestión del riesgo asegurador. 2014

200. El componente transfronterizo de las relaciones aseguradoras. 2014

199. El seguro basado en el uso (Usage Based Insurance). 2014

198. El seguro de decesos en la normativa aseguradora. Su encaje en Solvencia II. 2014

197. El seguro de responsabilidad civil en el arbitraje. 2014

196. La reputación corporativa en empresas aseguradoras: análisis y evaluación de factores explicativos. 2014

195. La acción directa del perjudicado en el ordenamiento jurídico comunitario. 2013

194. Investigaciones en Seguros y Gestión del Riesgo: RIESGO 2013

193. Viability of Patent Insurance in Spain. 2013

192. Viabilidad del seguro de patentes en España. 2013

191. Determinación de zonas homogéneas de riesgo para los rendimientos de distintos cultivos de la región pampeana en Argentina. 2013

190. Género y promoción en los sectores financiero y asegurador. 2013

189. An Introduction to Reinsurance. 2013

188. El control interno y la responsabilidad penal en la mediación de seguros privados. 2013

187. Una introducción al gobierno corporativo en la industria aseguradora en América Latina. 2013

186. Mortalidad de jóvenes en accidentes de tráfico. 2012

185. Las reclamaciones derivadas de accidentes de circulación por carretera transfronterizos. 2012

184. Efecto disuasorio del tipo de contrato sobre el fraude. 2012

183. Claves del Seguro Español: una aproximación a la Historia del Seguro en España. 2012

182. La responsabilidad civil del asegurador de asistencia sanitaria. 2012

181. Colaboración en el contrato de Reaseguro. 2012

180. Origen, situación actual y futuro del seguro de Protección Jurídica. 2012

179. Experiencias de microseguros en Colombia, Perú y Brasil. Modelo socio agente. 2012

178. El agente de seguros y su Responsabilidad Civil. 2012

177. Riesgo operacional en el marco de Solvencia II. 2012

176. Un siglo de seguros marítimos barceloneses en el comercio con América. (1770-1870). 2012

175. El seguro de Caución. 2012

174. La contabilidad de los corredores de seguros y los planes y fondos de pensiones. 2012

173. El seguro de Vida en América Latina. 2011

172. Gerencia de riesgos sostenibles y Responsabilidad Social Empresarial en la entidad aseguradora. 2011

171. Investigaciones en Seguros y Gestión del Riesgo. RIESGO 2011

170. Introdução ao Resseguro. 2011

169. La salud y su aseguramiento en Argentina, Chile, Colombia y España. 2011

168. Diferencias de sexo en conductas de riesgo y tasa de mortalidad diferencial entre hombres y mujeres. 2011

167. Movilización y rescate de los compromisos por pensiones garantizados mediante contrato de seguros. 2011

166. Embedded Value aplicado al ramo No Vida. 2011

165. Las sociedades cautivas de Reaseguro. 2011

164. Daños del amianto: litigación, aseguramiento de riesgos, y fondos de compensación. 2011

163. El riesgo de tipo de interés: experiencia española y Solvencia II. 2011

162. I Congreso sobre las Nuevas Tecnologías y sus repercusiones en el Seguro: Internet, Biotecnología y Nanotecnología. 2011

161. La incertidumbre bioactuarial en el riesgo de la longevidad. Reflexiones bioéticas. 2011

160. Actividad aseguradora y defensa de la competencia. La exención antitrust del sector asegurador. 2011

159. Estudio empírico sobre la tributación de los seguros de vida. 2010

158. Métodos estocásticos de estimación de las provisiones técnicas en el marco de Solvencia II. 2010

157. Introducción al Reaseguro. 2010

156. Encuentro Internacional sobre la Historia del Seguro. 2010

155. Los sistemas de salud en Latinoamérica y el papel del seguro privado. 2010

154. El Seguro de Crédito en Chile. 2010

153. El análisis financiero dinámico como herramienta para el desarrollo de modelos internos en el marco de Solvencia II. 2010

152. Características sociodemográficas de las personas con doble cobertura sanitaria. Un estudio empírico. 2010

151. Solidaridad impropia y seguro de Responsabilidad Civil. 2010

150. La prevención del blanqueo de capitales en las entidades aseguradoras, las gestoras y los corredores de seguros 2010

149. Fondos de aseguramiento agropecuario y rural: la experiencia mexicana en el mutualismo agropecuario y sus organizaciones superiores. 2010

148. Avaliação das Provisões de Sinistro sob o Enfoque das Novas Regras de Solvência do Brasil. 2010

147. El principio de igualdad sexual en el Seguro de Salud: análisis actuarial de su impacto y alcance. 2010

146. Investigaciones históricas sobre el Seguro español. 2010

145. Perspectivas y análisis económico de la futura reforma del sistema español de valoración del daño corporal. 2009

144. Contabilidad y Análisis de Cuentas Anuales de Entidades Aseguradoras (Plan contable 24 de julio de 2008). 2009

143. Mudanças Climáticas e Análise de Risco da Indústria de Petróleo no Litoral Brasileiro. 2009

142. Bases técnicas dinámicas del Seguro de Dependencia en España. Una aproximación en campo discreto. 2009

141. Transferencia Alternativa de Riesgos en el Seguro de Vida: Titulización de Riesgos Aseguradores. 2009

140. Riesgo de negocio ante asegurados con múltiples contratos. 2009

139. Optimización económica del Reaseguro cedido: modelos de decisión. 2009

138. Inversiones en el Seguro de Vida en la actualidad y perspectivas de futuro. 2009

137. El Seguro de Vida en España. Factores que influyen en su progreso. 2009

136. Investigaciones en Seguros y Gestión de Riesgos. RIESGO 2009

135. Análisis e interpretación de la gestión del fondo de maniobra en entidades aseguradoras de incendio y lucro cesante en grandes riesgos industriales. 2009

134. Gestión integral de Riesgos Corporativos como fuente de ventaja competitiva: cultura positiva del riesgo y reorganización estructural. 2009

133. La designación de la pareja de hecho como beneficiaria en los seguros de vida. 2009

132. Aproximación a la Responsabilidad Social de la empresa: reflexiones y propuesta de un modelo. 2009

131. La cobertura pública en el seguro de crédito a la exportación en España: cuestiones prácticas-jurídicas. 2009

130. La mediación en seguros privados: análisis de un complejo proceso de cambio legislativo. 2009

129. Temas relevantes del Derecho de Seguros contemporáneo. 2009

128. Cuestiones sobre la cláusula cut through. Transferencia y reconstrucción. 2008

127. La responsabilidad derivada de la utilización de organismos genéticamente modificados y la redistribución del riesgo a través del seguro. 2008

126. Ponencias de las Jornadas Internacionales sobre Catástrofes Naturales. 2008

125. La seguridad jurídica de las tecnologías de la información en el sector asegurador. 2008

124. Predicción de tablas de mortalidad dinámicas mediante un procedimiento bootstrap. 2008

123. Las compañías aseguradoras en los procesos penal y contencioso-administrativo. 2008

122. Factores de riesgo y cálculo de primas mediante técnicas de aprendizaje. 2008

121. La solicitud de seguro en la Ley 50/1980, de 8 de octubre, de Contrato de Seguro. 2008

120. Propuestas para un sistema de cobertura de enfermedades catastróficas en Argentina. 2008

119. Análisis del riesgo en seguros en el marco de Solvencia II: Técnicas estadísticas avanzadas Monte Carlo y Bootstrapping. 2008

118. Los planes de pensiones y los planes de previsión asegurados: su inclusión en el caudal hereditario. 2007

117. Evolução de resultados tecnicos e financieros no mercado segurador iberoamericano. 2007

116. Análisis de la Ley 26/2006 de Mediación de Seguros y Reaseguros Privados. 2007

115. Sistemas de cofinanciación de la dependencia: seguro privado frente a hipoteca inversa. 2007

114. El sector asegurador ante el cambio climático: riesgos y oportunidades. 2007

113. Responsabilidade social empresarial no mercado de seguros brasileiro influências culturais e implicações relacionais. 2007

112. Contabilidad y análisis de cuentas anuales de entidades aseguradoras. 2007

111. Fundamentos actuariales de primas y reservas de fianzas. 2007

110. El Fair Value de las provisiones técnicas de los seguros de Vida. 2007

109. El Seguro como instrumento de gestión de los M.E.R. (Materiales Específicados de Riesgo). 2006

108. Mercados de absorción de riesgos. 2006

107. La exteriorización de los compromisos por pensiones en la negociación colectiva. 2006

106. La utilización de datos médicos y genéticos en el ámbito de las compañías aseguradoras. 2006

105. Los seguros contra incendios forestales y su aplicación en Galicia. 2006

104. Fiscalidad del seguro en América Latina. 2006

103. Las NIC y su relación con el Plan Contable de Entidades Asegura-doras. 2006

102. Naturaleza jurídica del Seguro de Asistencia en Viaje. 2006

101. El Seguro de Automóviles en Iberoamérica. 2006

100. El nuevo perfil productivo y los seguros agropecuarios en Argentina. 2006

99. Modelos alternativos de transferencia y financiación de riesgos "ART": situación actual y perspectivas futuras. 2005

98. Disciplina de mercado en la industria de seguros en América Latina. 2005

97. Aplicación de métodos de inteligencia artificial para el análisis de la solvencia en entidades aseguradoras. 2005

96. El Sistema ABC-ABM: su aplicación en las entidades aseguradoras. 2005

95. Papel del docente universitario: ¿enseñar o ayudar a aprender? 2005

94. La renovación del Pacto de Toledo y la reforma del sistema de pensiones: ¿es suficiente el pacto político? 2005

92. Medición de la esperanza de vida residual según niveles de dependencia en España y costes de cuidados de larga duración. 2005

91. Problemática de la reforma de la Ley de Contrato de Seguro. 2005

90. Centros de atención telefónica del sector asegurador. 2005

89. Mercados aseguradores en el área mediterránea y cooperación para su desarrollo. 2005

88. Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación. 2004

87. Dependencia en el modelo individual, aplicación al riesgo de crédito. 2004

86. El margen de solvencia de las entidades aseguradoras en Iberoamérica. 2004

85. La matriz valor-fidelidad en el análisis de los asegurados en el ramo del automóvil. 2004

84. Estudio de la estructura de una cartera de pólizas y de la eficacia de un Bonus-Malus. 2004

83. La teoría del valor extremo: fundamentos y aplicación al seguro, ramo de responsabilidad civil autos. 2004

81. El Seguro de Dependencia: una visión general. 2004

80. Los planes y fondos de pensiones en el contexto europeo: la necesidad de una armonización. 2004

79. La actividad de las compañías aseguradoras de vida en el marco de la gestión integral de activos y pasivos. 2003

78. Nuevas perspectivas de la educación universitaria a distancia. 2003

77. El coste de los riesgos en la empresa española: 2001.

76. La incorporación de los sistemas privados de pensiones en las pequeñas y medianas empresas. 2003

75. Incidencia de la nueva Ley de Enjuiciamiento Civil en los procesos de responsabilidad civil derivada del uso de vehículos a motor. 2002

74. Estructuras de propiedad, organización y canales de distribución de las empresas aseguradoras en el mercado español. 2002

73. Financiación del capital-riesgo mediante el seguro. 2002

72. Análisis del proceso de exteriorización de los compromisos por pensiones. 2002

71. Gestión de activos y pasivos en la cartera de un fondo de pensiones. 2002

70. El cuadro de mando integral para las entidades aseguradoras. 2002

69. Provisiones para prestaciones a la luz del Reglamento de Ordenación y Supervisión de los Seguros Privados; métodos estadísticos de cálculo. 2002

68. Los seguros de crédito y de caución en Iberoamérica. 2001

67. Gestión directiva en la internacionalización de la empresa. 2001

65. Ética empresarial y globalización. 2001

64. Fundamentos técnicos de la regulación del margen de solvencia. 2001

63. Análisis de la repercusión fiscal del seguro de vida y los planes de pensiones. Instrumentos de previsión social individual y empresarial. 2001

62. Seguridad Social: temas generales y régimen de clases pasivas del Estado. 2001

61. Sistemas Bonus-Malus generalizados con inclusión de los costes de los siniestros. 2001

60. Análisis técnico y económico del conjunto de las empresas aseguradoras de la Unión Europea. 2001

59. Estudio sobre el euro y el seguro. 2000

58. Problemática contable de las operaciones de reaseguro. 2000

56. Análisis económico y estadístico de los factores determinantes de la demanda de los seguros privados en España. 2000

54. El corredor de reaseguros y su legislación específica en América y Europa. 2000

53. Habilidades directivas: estudio de sesgo de género en instrumentos de evaluación. 2000

52. La estructura financiera de las entidades de seguros, S.A. 2000

51. Seguridades y riesgos del joven en los grupos de edad. 2000

50. Mixturas de distribuciones: aplicación a las variables más relevantes que modelan la siniestralidad en la empresa aseguradora. 1999

49. Solvencia y estabilidad financiera en la empresa de seguros: metodología y evaluación empírica mediante análisis multivariante. 1999

48. Matemática Actuarial no vida con MapleV. 1999

47. El fraude en el Seguro de Automóvil: cómo detectarlo. 1999

46. Evolución y predicción de las tablas de mortalidad dinámicas para la población española. 1999

45. Los Impuestos en una economía global. 1999

42. La Responsabilidad Civil por contaminación del entorno y su aseguramiento. 1998

41. De Maastricht a Amsterdam: un paso más en la integración europea. 1998

39. Perspectiva histórica de los documentos estadístico-contables del órgano de control: aspectos jurídicos, formalización y explotación. 1997

38. Legislación y estadísticas del mercado de seguros en la comunidad iberoamericana. 1997

37. La responsabilidad civil por accidente de circulación. Puntual comparación de los derechos francés y español. 1997

36. Cláusulas limitativas de los derechos de los asegurados y cláusulas delimitadoras del riesgo cubierto: las cláusulas de limitación temporal de la cobertura en el Seguro de Responsabilidad Civil. 1997

35. El control de riesgos en fraudes informáticos. 1997

34. El coste de los riesgos en la empresa española: 1995

33. La función del derecho en la economía. 1997

32. Decisiones racionales en reaseguro. 1996

31. Tipos estratégicos, orientación al mercado y resultados económicos: análisis empírico del sector asegurador español. 1996

30. El tiempo del directivo. 1996

29. Ruina y Seguro de Responsabilidad Civil Decenal. 1996

28. La naturaleza jurídica del Seguro de Responsabilidad Civil. 1995

27. La calidad total como factor para elevar la cuota de mercado en empresas de seguros. 1995

26. El coste de los riesgos en la empresa española: 1993

25. El reaseguro financiero. 1995

24. El seguro: expresión de solidaridad desde la perspectiva del derecho. 1995

23. Análisis de la demanda del seguro sanitario privado. 1993

22. Rentabilidad y productividad de entidades aseguradoras. 1994

21. La nueva regulación de las provisiones técnicas en la Directiva de Cuentas de la C.E.E. 1994

20. El Reaseguro en los procesos de integración económica. 1994

19. Una teoría de la educación. 1994

18. El Seguro de Crédito a la exportación en los países de la OCDE (evaluación de los resultados de los aseguradores públicos). 1994

16. La legislación española de seguros y su adaptación a la normativa comunitaria. 1993

15. El coste de los riesgos en la empresa española: 1991

14. El Reaseguro de exceso de pérdidas 1993

12. Los seguros de salud y la sanidad privada. 1993

10. Desarrollo directivo: una inversión estratégica. 1992

9. Técnicas de trabajo intelectual. 1992

8. La implantación de un sistema de controlling estratégico en la empresa. 1992

7. Los seguros de responsabilidad civil y su obligatoriedad de aseguramiento. 1992

6. Elementos de dirección estratégica de la empresa. 1992

5. La distribución comercial del seguro: sus estrategias y riesgos. 1991

4. Los seguros en una Europa cambiante: 1990-95. 1991

2. Resultados de la encuesta sobre la formación superior para los profesionales de entidades aseguradoras (A.P.S.). 1991

1. Filosofía empresarial: selección de artículos y ejemplos prácticos. 1991

ÚLTIMOS LIBROS PUBLICADOS

El riesgo de longevidad y su aplicación práctica a Solvencia II. 2014

Historia del Seguro en España. 2014

Actas del III Congreso Internacional de Nuevas Tecnologías: sus repercusiones en el seguro: internet, biotecnología y nanotecnología: 12 y 13 de noviembre de 2012, Santiago de Chile. 2013

Emergencia y reconstrucción: el antes y después del terremoto y tsunami del 27F en Chile. 2012

Riesgo sistémico y actividad aseguradora. 2012

La historia del seguro en Chile (1810-2010). 2012

Modelo de proyección de carteras de seguros para el ramo de decesos. 2011

Desarrollo comercial del seguro colectivo de dependencia en España. 2010

La mediación de seguros en España: análisis de la Ley 26/2006, de Mediación de Seguros y Reaseguros Privados. 2010

Museo del Seguro. Catálogo. 2010

Diccionario MAPFRE de Seguros. 2008

Teoría de la credibilidad: desarrollo y aplicaciones en primas de seguros y riesgos operacionales. 2008

El seguro de caución: una aproximación práctica. 2007

El seguro de pensiones. 2007

Las cargas del acreedor en el seguro de responsabilidad civil. 2006

Diccionario bilingüe de expresiones y términos de seguros: inglés-español, español-inglés. 2006

El seguro de riesgos catastróficos: reaseguro tradicional y transferencia alternativa de riesgos. 2005

La liquidación administrativa de entidades aseguradoras. 2005

INFORMES Y RANKINGS

Desde 1994 se publican anualmente estudios que presentan una panorámica concreta de los mercados aseguradores europeos, de España e Iberoamérica y que pueden consultarse en formato electrónico desde la página Web: www.fundacionmapfre.org

- Mercado español de seguros
- Mercado asegurador de Iberoamérica
- Ranking de grupos aseguradores europeos
- Ranking de grupos aseguradores iberoamericanos
- La percepción social del seguro en España 2014
- Informe de predicción de la actividad aseguradora en España. 2014
- La internacionalización de la empresa española: riesgos y oportunidades. 2014
- El seguro en la sociedad y la economía españolas. 2013
- Papel del seguro en el desarrollo sostenible. ICEA, 2013
- Emprender en momentos de crisis: riesgos y factores de éxito. 2012
- La percepción social del seguro en España 2012

FUNDACIÓNMAPFRE

2015

Paseo de Recoletos, 23
28004 Madrid (España)
www.fundacionmapfre.org

P.V.P.: 35 €
ISBN 978-84-9844-496-4