TÉCNICAS CUANTITATIVAS PARA LA DETECCIÓN DEL FRAUDE EN EL SEGURO DEL AUTOMÓVIL

Mercedes Ayuso, Montserrat Guillén y Manuel Artís
Grupo de investigación: Riesgo en Finanzas y Seguros
Departamento de Econometría, Estadística y Economía Española
Universidad de Barcelona

RESUMEN

La aplicación de técnicas estadísticas y econométricas dirigidas a cuantificar la probabilidad de fraude en los expedientes de siniestros automovilísticos está ganando terreno en los últimos años. La validación estadística de los denominados *indicadores de fraude* es, sin duda, una pieza clave a la hora de dirigir de forma adecuada la investigación de los accidentes.

Los estudios de carácter cuantitativo sobre el fraude no son numerosos. La literatura existente presenta, fundamentalmente, un enfoque teórico-económico teniendo como principal objetivo modelizar cómo influye la existencia de información asimétrica en la formalización y posterior aplicación del contrato de seguro. Adicionalmente, es posible encontrar una serie de manuales, muchas veces editados por las propias compañías, dirigidos a motivar al personal asegurador en la lucha contra el problema.

Sin embargo, las posibilidades que la estadística y la econometría ofrecen para realizar estudios de carácter aplicado son numerosas. En este artículo se presenta una revisión de las principales contribuciones existentes, fundamentadas, todas ellas, en el uso de muestras de expedientes de siniestros reales.

PALABRAS CLAVE: modelos cuantitativos, seguro del automóvil, indicadores de fraude, tipología de fraude, regresión logística anidada.

INTRODUCCIÓN

La existencia de comportamientos fraudulentos en el seguro del automóvil es un hecho aceptado dentro de la comunidad aseguradora. Considerar el fraude como un factor ineludible de riesgo ha perdido vigencia y, hoy en día, las entidades luchan por desarrollar un foco de acción frente al mismo. La influencia de las acciones deshonestas por parte de los asegurados se deja sentir tanto en el número de siniestros declarados como en la cuantía de los mismos. Si consideramos el peso que ello puede tener a la hora de justificar la aparición de resultados técnicos negativos durante los últimos años en el seguro del automóvil, queda más que justificada la necesidad de diseñar herramientas que ayuden a las entidades en la lucha contra el fraude.

Artículos diversos publicados en revistas especializadas del sector e incluso manuales como la *Guía Anti-fraude*, editada por el Comité Europeo de Seguros[1] (1996) o el libro, *Manual de Investigación de Siniestros y Lucha contra el Fraude en el Seguro de Automóviles* (Cobo, 1993) alertan sobre la existencia del problema, sobre sus diferentes formas de manifestarse y señalan pautas de actuación frente al mismo. En todos ellos se hace referencia a la importancia de diseñar un perfil del asegurado defraudador y a la conveniencia de identificar determinados factores que incrementan la probabilidad de aparición de fraude. A pesar de ello, el diseño de métodos cuantitativos dirigidos al estudio de dichos factores o de las circunstancias comúnmente asociadas a la aparición de comportamientos deshonestos ha sido hasta hace muy poco tiempo inexistente, no sólo dentro del marco nacional, sino también del internacional.

La alarma desencadenada dentro del sector asegurador en torno al problema y, la posibilidad de tratar el mismo desde un punto de vista aplicado, ha despertado el interés de los investigadores. Los últimos años han supuesto un giro de 90º y si hasta hace poco tiempo la literatura existente sobre el tema era prácticamente inexistente, hoy en día es posible encontrar estudios, cada vez más exhaustivos, dirigidos

[1] *Le Guide de l'anti-fraude à l'assurance en Europe*.

a validar econométricamente los denominados "indicadores de fraude".

Según manifestó el Dr. Richard Derrig, vicepresidente del Insurance Fraud Bureau de Massachusetts (en adelante, I.F.B.), en una conferencia sobre el fraude en el seguro del automóvil celebrada en Barcelona en abril de 1998, es posible hablar de la existencia de tres grupos de investigación que, a nivel mundial, modelizan el comportamiento deshonesto dentro de las reclamaciones de siniestros: el suyo propio; el creado en la École des Hautes Études Commerciales de la Universidad de Montreal (al que pertenecen autores de reconocido prestigio dentro del campo científico asegurador como Georges Dionne) y el creado en el Departamento de Econometría, Estadística y Economía Española de la Universidad de Barcelona.

Como veremos a lo largo del presente artículo, no se encuentran diferencias notables (o que no puedan explicarse por los distintos sistemas aseguradores) entre los resultados obtenidos por los tres grupos de trabajo. Las divergencias se dejan sentir, principalmente, en el tratamiento metodológico aplicado y en la definición utilizada de fraude. El objetivo perseguido es claro: detectar y en la medida de lo posible, sistematizar, señales de alerta que ayuden al personal de la compañía a realizar una correcta política de suscripción de pólizas y de tramitación de siniestros. En definitiva, realizar una primera aproximación a lo que podría llegar a constituirse en una política de control y detección de fraude en el seguro.

Aunque cada vez está ganando una mayor importancia en otros ramos, la aplicación de métodos cuantitativos estadísticos y econométricos al tratamiento del fraude se ha centrado, fundamentalmente, en el seguro del automóvil. A lo largo del presente artículo realizaremos un análisis de las principales aportaciones realizadas en dicho contexto asegurador, teniendo en cuenta diferentes vertientes metodológicas.

El orden de exposición será el siguiente. En primer lugar presentaremos un resumen de las diferentes técnicas estadísticas y econométricas propuestas para modelizar el fraude en el seguro automovilístico. En segundo lugar, especificaremos los modelos de

regresión logística anidada. En tercer y, último lugar, analizaremos los principales resultados obtenidos en una muestra de expedientes de siniestros del mercado asegurador español.

1. TÉCNICAS CUANTITATIVAS DE CONTROL Y DETECCIÓN DEL FRAUDE.

La modelización del comportamiento deshonesto de los asegurados dentro del seguro del automóvil ha dado lugar a la aplicación de métodos cuantitativos diversos. Aunque muchos de los trabajos existentes han sido calificados por los propios autores de preliminares, todos ellos señalan la posibilidad de dar diferentes matices al tratamiento del problema. Los resultados obtenidos revelan, en la mayoría de los casos, la existencia de coincidencias para determinadas variables o indicadores que siempre irán asociadas a un aumento o disminución en la probabilidad de aparición de fraude[2].

El reducido tamaño de las bases de datos empleadas en algunos estudios realizados hasta el momento hace que las conclusiones extraídas deban tomarse con la debida precaución. El desarrollo de un análisis exhaustivo del fraude conlleva una dificultad añadida: la disponibilidad de muestras de expedientes de siniestros lo suficientemente grandes como para poder extrapolar los resultados obtenidos al conjunto poblacional. Como veremos a continuación, las muestras utilizadas en los diferentes estudios son de tamaño reducido y recogen, normalmente, información para siniestros con y sin sospecha de fraude. Sólo en el trabajo realizado para un segmento del mercado español se ha podido disponer de una muestra de expedientes con fraude detectado.

En muchos casos la reticencia de las compañías a facilitar información ha ido acompañada de la dificultad que genera el diseño correcto de una base de datos para hacerla susceptible de tratamiento estadístico. Además, la reducida exhaustividad que suele rodear la recogida de datos en el parte de siniestro (al menos en nuestro país es frecuente

[2] A modo de ejemplo, la intervención de policía en el siniestro reducirá normalmente el espacio que el asegurado posee para cometer fraude.

encontrar declaraciones de accidente con gran número de espacios en blanco) disminuye el número de variables susceptibles de ser consideradas en la modelización. Sólo en Massachusetts, la elaboración de una gran base de datos, conocida como la *D.C.D.*[3]*,* regulada legalmente y con información para todos los siniestros cerrados en o a partir del 1 de enero de 1994 por las entidades aseguradoras de dicho estado, está permitiendo al I.F.B.[4] trabajar con volúmenes de datos más elevados.

La validación estadística de los indicadores de fraude que deben ser considerados por las entidades de cara a identificar un suceso como fraudulento se ha realizado, fundamentalmente, considerando cuatro vertientes metodológicas:

- la especificación de modelos de regresión lineal múltiple;
- el uso de la teoría de conjuntos borrosos;
- la aplicación de redes neuronales y,
- el planteamiento de modelos de elección discreta.

Nuestro trabajo, enmarcado en la última de las vertientes presentadas, persigue cuantificar la probabilidad de que un individuo decida comportarse honesta o fraudulentamente atendiendo a determinadas características del mismo y también de la póliza, del vehículo y del siniestro. La selección de las variables acompañadas de coeficientes estadísticamente significativos ("indicadores de fraude") y la posibilidad de implementar informáticamente el modelo especificado supondrá para la entidad un primer paso de cara a diferenciar de forma ágil entre siniestros con y sin sospecha de fraude.

[3] *Detailed Claim Database*.
[4] El control de la D.C.D. lo realiza el A*utomobile Insurers Bureau* de Massachusetts, dentro del cual se integra el *Insurance Fraud Bureau*.

2. APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE.

El modelo de regresión lineal múltiple, técnica econométrica clásica, ha sido la herramienta básica utilizada en los trabajos realizados desde el I.F.B. de Massachusetts.

El objetivo perseguido por este organismo ha sido determinar aquellos indicadores sobre los que la compañía ha de realizar una investigación exhaustiva de cara a verificar la existencia de comportamiento fraudulento.

Atendiendo a la definición de fraude, diferencian cuatro situaciones posibles a la hora de clasificar un siniestro[5] (Weisberg y Derrig, 1993):

- legítimo (sin fraude),
- fraude oportunista (casos en los que ocurre el siniestro y el asegurado oportunista lo aprovecha para declarar un daño inexistente),
- aumento deliberado de los gastos médicos derivados del siniestro (existe el siniestro, existe daño, pero el asegurado exagera los gastos que se derivan del mismo) y,
- fraude planeado (situaciones en las que la ocurrencia del siniestro no es accidental sino que es provocada deliberadamente).

En trabajos previos (Weisberg y Derrig, 1991), teniendo en cuenta la explotación de una muestra de 597 siniestros de autos ocurridos en Massachusetts entre 1985 y 1986 ("*The Baseline Study"*), se había confirmado la asociación existente entre la presencia de fraude y la declaración de determinados daños corporales (se observaba un mayor número de torceduras y daños sin importancia considerable en los siniestros fraudulentos).

[5] En terminología americana se habla de *legitimate claim*, *opportunistic fraud*, *build-up* and *planned fraud*.

Sin embargo, es en 1993 cuando Weisberg y Derrig presentan los resultados derivados de la aplicación de la técnica de regresión lineal múltiple a una base de datos más elaborada. En este caso, la muestra estaba constituida por un total de 387 expedientes de siniestros acaecidos en 1989, también en Massachusetts. No obstante, cuestiones prácticas llevaron a realizar la estimación sólo con una parte de la misma (del total de la muestra sólo en 62 expedientes se detectó una elevada sospecha de fraude; la submuestra de trabajo fue enriquecida con 65 expedientes seleccionados al azar entre los 325 restantes).

Para los 127 expedientes de siniestros finales se disponía de información de 65 indicadores de fraude, relacionados con diferentes aspectos del siniestro y de las partes involucradas en el mismo. Cada uno de los expedientes analizados fue objeto de valoración en términos de sospecha de fraude en una escala de 0 a 10. En dicho proceso, el analista[6] tuvo en cuenta el nivel de sospecha que le infundía el propio accidente, el demandante, el asegurado, el daño, el tratamiento médico y la pérdida de salario derivada. Asimismo, para cada uno de los expedientes se recogió en una variable el número de votos a favor de la existencia de fraude (teniendo en cuenta la opinión de tramitadores e investigadores).

En la especificación de los modelos se perseguían, básicamente, dos objetivos. En un caso, la variable a explicar era una variable cuantitativa que recogía el nivel de sospecha de fraude en los expedientes de siniestros analizados (se presentaban por separado los modelos propuestos por los tramitadores profesionales y los propuestos por los investigadores independientes), modelizándose, de forma conjunta, los cuatro tipos de comportamiento comentados anteriormente. En el otro, dicha variable, también cuantitativa, recogía el número de votos a favor de la existencia de fraude (sin diferenciar el tipo).

[6] Cada expediente fue objeto de análisis por un grupo de tramitadores profesionales y por un grupo de investigadores especializados del Insurance Fraud Bureau.

La validación global de los modelos propuestos se realizó a partir del coeficiente de determinación o bondad del ajuste[7]. El modelo con variable dependiente el índice de sospecha de fraude para los tramitadores presentaba una bondad del ajuste del 65%, coeficiente que disminuía hasta el 56% cuando se consideraba el índice de sospecha de fraude para los investigadores. En el modelo con variable dependiente el número de votos a favor de la presencia de fraude, la bondad del ajuste fue sólo del 46%.

En la actualidad, el objetivo perseguido bajo el denominado *Claim Screening Experiment, C.S.E.*(Derrig y Weisberg, 1998), es modelizar la sospecha de fraude en diferentes etapas de la "vida" del siniestro. Una vez ocurrido el accidente y declarado a la compañía, la información relacionada con diferentes aspectos del mismo (fundamentalmente, si se producen daños a la persona con el consiguiente tratamiento médico) llegará una vez han transcurrido diferentes periodos de tiempo. La idea es diseñar un método de control que alerte al tramitador sobre la conveniencia de investigar o no un siniestro a medida que se va recibiendo información.

Las técnicas de regresión clásica se han aplicado también en estudios de fraude realizados desde una óptica macroeconómica o agregada. Cummins y Tennyson (1996) presentan un enfoque sectorial introduciendo como variables explicativas indicadores relacionados con el sistema económico. Los autores modelizan la presencia de fraude a nivel agregado de un estado considerando su incidencia en la frecuencia de siniestros declarados por daños corporales. Para ello estiman un modelo lineal para el logaritmo del cociente entre la frecuencia de siniestros por daños corporales y la frecuencia por daños materiales. La muestra utilizada está formada por observaciones recogidas en 29 estados norteamericanos en 1991 y 1992. Se concluye que diversas actitudes de los individuos hacia el fraude (permitir al médico extender facturas por servicios no prestados, mentir sobre las pérdidas para recuperar las primas pagadas,...) pueden considerarse indicadores de riesgo moral.

[7] Las medidas de calidad del ajuste son aquellas destinadas a evaluar en que medida el modelo utilizado explica las variaciones que se producen en la variable dependiente.

3. APLICACIÓN DE LA TEORÍA DE CONJUNTOS BORROSOS.

En la década de los noventa, la aplicación de la teoría de conjuntos borrosos en el campo financiero y asegurador dio lugar a trabajos variados. Derrig y Ostaszewski (1995) presentan, por primera vez, una aplicación de las técnicas de conjuntos borrosos a la clasificación de siniestros en términos de sospecha de fraude. Utilizan esta técnica para realizar una agrupación de expedientes de siniestros teniendo en cuenta la existencia de diferentes tipos de fraude. Para realizar su estudio, emplean la misma muestra que había sido utilizada en estudios previos para Massachusetts y que hemos comentado en el apartado anterior, formada por 127 expedientes de siniestros.

La modelización del nivel de sospecha de fraude globalmente (dando valores a la variable dependiente en una escala de 0 a 10) no permite diferenciar qué indicadores tienen mayor peso a la hora de explicar la aparición de diferentes tipos de fraude. El intento de generar diferentes subgrupos (legítimo, fraude oportunista, siniestro con costes médicos hinchados y fraude planeado) considerando el nivel de sospecha medio (Weisberg y Derrig, 1993) para cada una de las categorías de clasificación puede llevar a conclusiones erróneas, sobre todo si consideramos que la diferencia entre los distintos comportamientos se hace patente al analizar, por separado, el nivel de sospecha para cada uno de los aspectos o componentes del siniestro. De hecho, cuando el siniestro es legítimo el grado de sospecha es bajo para todos los componentes analizados (accidente, asegurado, daño, tratamiento,...). Cuando el siniestro es planeado la sospecha es elevada en prácticamente todas las vertientes analizadas[8]. Ahora bien cuando hablamos de fraude oportunista y de hinchamiento de costes médicos, la sospecha será elevada en determinados componentes y baja en otros.

La aplicación de la teoría de conjuntos borrosos permite sustituir la dicotomía fraude-no fraude por una función de medida cuyos valores

[8] Únicamente para la "pérdida de salario derivada del accidente" el nivel de sospecha está muy por debajo de siete. La razón se encuentra en que, normalmente, este tipo de fraude lo cometen personas desempleadas.

oscilan, precisamente, entre dichos extremos. Si una vez realizado el análisis se obtiene una valoración cercana a cero o a uno la clasificación del expediente está clara. Ahora bien, si se obtiene un valor que se encuentra a lo largo del intervalo considerado, la técnica generará una clasificación determinada en función de la sospecha de fraude existente. Además, permitirá observar la existencia de categorías de clasificación muy similares.

Atendiendo a la sospecha de fraude en una escala de cero a diez, las respuestas ofrecidas por los tramitadores de siniestros para los niveles de sospecha se clasifican en cinco grupos: ausencia de sospecha (0), sospecha débil (1-3), sospecha moderada (4-6), sospecha fuerte (7-9) y sospecha cierta (10). Finalmente, el algoritmo utilizado genera una clasificación final de los siniestros, poniéndose de manifiesto, a través de las diferencias entre los centros de los diferentes grupos, una divergencia entre el enfoque utilizado por tramitadores e investigadores. Esta divergencia es especialmente importante a la hora de clasificar el siniestro de fraude oportunista o de hinchamiento de costes médicos. El elevado nivel de sospecha declarado para las componentes que recogen los daños a la persona y el tratamiento médico en ambos tipos de comportamiento dificulta su diferenciación.

Para realizar la clasificación en las diferentes categorías de fraude, cada siniestro es objeto de codificación teniendo en cuenta el nivel de sospecha existente (de nuevo escalado de cero a diez) en las seis componentes consideradas (el accidente, el daño, el asegurado, el demandante, el tratamiento médico y la pérdida de salario). Finalmente, los resultados obtenidos muestran una agrupación en cinco tipos de comportamiento (se produce un desdoblamiento en la categoría de "hinchamiento de costes médicos"). El centro de cada uno de ellos señala, considerando el grado de sospecha para cada una de las seis componentes analizadas, la posible presencia de uno u otro tipo de fraude. A modo de ejemplo, el cluster con centro (0,0,0,0,0,0) recoge siniestros sin sospecha de fraude mientras que en el cluster con centro (7,8,7,8,8,0) aparecen siniestros con alta sospecha de fraude planeado.

En base al sistema generado, la valoración realizada para las diferentes componentes del siniestro permitirá clasificar al siniestro dentro de una u otra categoría. Sin embargo los resultados obtenidos podrían llevar asociado un cierto matiz de subjetividad dado que parten de valoraciones realizadas por tramitadores e investigadores para el nivel de sospecha en relación a las componentes analizadas del siniestro. La agrupación de siniestros sospechosos podría realizarse considerando datos multidimensionales relacionados, no sólo con valoraciones subjetivas, sino con comportamientos reales observados para los diferentes aspectos o circunstancias que rodean el acaecimiento del suceso (fundamentalmente, para aquellos que pueden considerarse posibles indicadores de fraude).

4. APLICACIÓN DE REDES NEURONALES.

En 1995, Brockett, Xia y Derrig muestran la posibilidad de construir un sistema de detección de fraude teniendo en cuenta indicadores objetivos y subjetivos de fraude, utilizando redes neuronales.

En su estudio utilizan, de nuevo, la muestra no aleatoria formada por 127 expedientes de siniestros que había sido usada en estudios anteriores, obteniendo información para el conjunto de 65 indicadores que ya habían sido considerados por Weisberg y Derrig (1993).

Dado que uno de los objetivos perseguidos es especificar un modelo que posea elevada calidad predictiva (el error a la hora de clasificar un nuevo siniestro no considerado en la muestra de estudio ha de ser el menor posible), los autores seleccionan 77 siniestros del total de la muestra para realizar la modelización y dejan los 50 restantes para predecir *ex-post*[9]. Partiendo de las premisas de que modelos de siniestros parecidos presentarán niveles de sospecha similares y de que cada uno de los indicadores considerados tendrá igual importancia a la hora de explicar la existencia de fraude, la aplicación de redes neuronales supone la representación de cada uno de los siniestros mediante un vector de atributos. En este caso dicho vector está

[9] Se utiliza una parte de la muestra para validar los resultados obtenidos.

compuesto por los 65 indicadores de fraude para los que se dispone de información.

Teniendo en cuenta un *input* formado por 77 vectores (uno para cada siniestro utilizado en la modelización), el *output* resultante se centra en la clasificación del suceso en una de las siguientes categorías: siniestro válido o legal, siniestro con débil sospecha de fraude, siniestro con sospecha de fraude moderada y siniestro con elevada sospecha de fraude. El hecho de que dentro del mapa de representación, los siniestros con vectores de indicadores parecidos y, por tanto, con distancias pequeñas entre ellos, queden más o menos juntos y alejados del resto, permite crear zonas, *regiones de decisión*, que se identifican con las cuatro categorías de clasificación comentadas.

Las conclusiones obtenidas del estudio revelan una elevada calidad del modelo en términos de predicción *ex-post*, sin que pueda decirse lo mismo en relación a la predicción *ex-ante* o predicción propiamente dicha (para siniestros no considerados en la muestra), aunque quizá una explicación pueda encontrarse en el reducido tamaño muestral. El hecho de que todos los indicadores sean ponderados por igual impide determinar la existencia de variables con mayor poder explicativo que otras.

5. APLICACIÓN DE MODELOS DE ELECCIÓN PROBABILÍSTICA.

Los modelos de elección discreta permiten cuantificar la probabilidad de aparición de comportamientos deshonestos cuando la variable dependiente ha sido adecuadamente categorizada para recoger la dicotomía presencia/ausencia de fraude. En este caso, la variable a explicar deja de ser una variable cuantitativa e indica mediante una determinada codificación (normalmente, uno-cero) el cumplimiento o no de una característica dada. Lógicamente, dicha codificación variará en el caso de trabajar con múltiples categorías de elección.

La aplicación de este tipo de modelos al análisis del fraude en el seguro del automóvil ha dado lugar a estudios diversos.

Belhadji y Dionne (1997) plantean un modelo próbit simple para estimar la probabilidad de existencia de fraude (frente a la de no fraude), en una muestra de 2068 expedientes de siniestros del mercado asegurador automovilístico de Quebec (Canadá). Asimismo, en Ayuso y Guillén[10] (1995a, 1999) se puede encontrar una modelización de la dicotomía planteada mediante la aplicación de modelos logit simples a una muestra de expedientes de siniestros norteamericana y española, respectivamente.

Ahora bien, ¿qué ocurre cuando pretendemos modelizar la existencia de comportamiento fraudulento teniendo en cuenta las diferentes formas de manifestarse?. En este caso, los modelos logit y próbit simples no son susceptibles de aplicación y se hace necesario el uso de una modelización econométrica más sofisticada. Artís, Ayuso y Guillén (1999) presentan los resultados de aplicar modelos logísticos multinomiales y anidados a una muestra de expedientes del mercado asegurador español. El objetivo, como detallaremos más adelante, es considerar en la modelización el proceso de decisión que lleva al asegurado a comportarse honestamente o a cometer un determinado tipo de fraude.

5.1 MODELOS LOGIT Y PROBIT SIMPLES.

Ayuso y Guillén (1995a) aplican un modelo de regresión logística simple para cuantificar la probabilidad de existencia de fraude con la muestra utilizada en el estudio de Massachusetts.

[10] En el artículo "Modelos de detección de Fraude en el Seguro del Automóvil" (Ayuso y Guillén, 1999), considerando los resultados obtenidos en trabajos anteriores, se sugiere la conveniencia de diferenciar entre fraude "*a priori*" y fraude *"a posteriori*". Bajo la primera denominación se recogen todas aquellas situaciones en las que la ocurrencia del siniestro ha sido planeada; bajo la segunda, aquellas en las que el siniestro realmente ocurre pero el asegurado lo aprovecha para actuar en beneficio propio.

Considerando la tipología de fraude propuesta en EE.UU., el objetivo del estudio se centró en modelizar, por separado, tres situaciones alternativas: a) ausencia de fraude *versus* existencia de fraude; b) ausencia de fraude *versus* existencia de fraude "*a priori*[11] " y, c) ausencia de fraude *versus* existencia de fraude "*a posteriori*[12]". Las variables explicativas introducidas en la especificación, relacionadas con diferentes aspectos del siniestro, de la póliza y del demandante, recogían en algunos casos información que no suele ser objeto de estudio dentro del mercado asegurador español (por ejemplo, la situación laboral del demandante).

En aquel momento, los resultados obtenidos de la modelización (satisfactorios tanto en términos de significación individual y global de los modelos como en capacidad predictiva) permitieron diferenciar el trabajo del desarrollado simultáneamente por otros autores en un aspecto fundamental. El objetivo perseguido no era tanto determinar conjuntos de variables a las que acompañasen coeficientes estadísticamente significativos (sobre las que la entidad dirigiría la investigación) como cuantificar la probabilidad de aparición de fraude en el análisis de una reclamación. Lógicamente, este proceso permitiría evaluar la importancia relativa de las variables explicativas del fraude y si su influencia era significativamente diferente de cero.

La modelización, por separado, de los diferentes tipos de comportamientos fraudulentos reveló la existencia de variables con influencia muy distinta en la aparición de las diferentes clases de fraude. A modo de ejemplo, el hecho de que el demandante no tuviera trabajo era una variable que únicamente poseía influencia en la aparición de fraude *a priori*.

Siguiendo un proceso similar pero, sin tener en cuenta la existencia de diferentes tipos de comportamientos fraudulentos, Belhadji y Dionne (1997) especifican un modelo próbit sencillo para cuantificar la probabilidad de existencia de fraude (detectado o sospechado).

[11] En este caso la muestra estaba formada por los siniestros considerados legítimos y aquellos en los que se sospechaba de la existencia de fraude planeado.

[12] Siniestros legítimos frente a siniestros con sospecha de fraude oportunista y siniestros con sospecha de hinchamiento de costes médicos.

Los objetivos perseguidos por ambos autores son fundamentalmente dos: el diseño de un mecanismo de detección basado en la implementación automática del modelo próbit especificado (dirigido a cuantificar la probabilidad de que el siniestro sea fraudulento) y el desarrollo de un sistema que analice la conveniencia o no de investigar los siniestros con elevada sospecha de fraude teniendo en cuenta la relación coste beneficio[13].

Como resultado de la estimación, 18 variables aparecen acompañadas de coeficientes estadísticamente significativos. Entre ellas destaca el hecho de que el asegurado esté extraordinariamente familiarizado con el proceso de tramitación y la jerga empleada en seguros y reparaciones, el que posea dificultades financieras personales o de su negocio, el que durante la investigación parezca nervioso y confuso o el que reclame un elevado número de facturas médicas. Al igual que Weisberg y Derrig (1993), los autores introducen en la modelización variables subjetivas, generadas en base a valoraciones realizadas por el tramitador del siniestro (grado de nerviosismo del declarante, agresividad del mismo,...). También consideran variables para las que la entidad dispone de información de manera progresiva.

La insuficiencia de datos hizo que el mecanismo diseñado por ambos autores para evaluar la conveniencia o no de investigar el siniestro, considerando la relación coste beneficio, fuese preliminar. En su aproximación hacen depender el coste de la investigación de cuatro factores: la probabilidad de fraude, la experiencia de los investigadores, su formación y la existencia de unidades especiales de investigación dentro de la compañía.

[13] En este caso, la muestra utilizada recogía información para 18 compañías del mercado asegurador automovilístico de Quebec. De los 2068 expedientes que la componían, 1937 habían sido clasificados como no fraudulentos, 113 contenían sospecha de fraude y en 18 existía fraude detectado. La selección se realizó aleatoriamente por las entidades (cada compañía participó de forma proporcional a su cuota de mercado) entre todos los expedientes cerrados entre el 1 de abril de 1994 y el 31 de marzo de 1995.

El trabajar con una muestra de expedientes con fraude detectado, el disponer de variables para las que la entidad dispone de información de manera inmediata y el hecho de que éstas posean carácter objetivo y no dependan de valoraciones personales, son, como veremos a continuación, las principales diferencias de nuestra propuesta.

5.2 MODELOS LOGIT MULTINOMIALES Y ANIDADOS.

La aproximación que ahora presentamos tiene en cuenta, en la modelización, la tipología existente de fraude dentro del seguro automovilístico español. El falseamiento de la declaración para eludir casos excluidos en la póliza, la ocultación de circunstancias personales del asegurado o la falsa declaración para favorecer a un tercero, aparecen como algunos de los comportamientos fraudulentos más frecuentes en España (ICEA, 2000) y es lógico pensar en la existencia de características diferenciales en torno a los mismos. Además, trabajamos bajo la hipótesis de que la diferenciación en clases de fraude es imprescindible para el diseño de una eficaz política/estrategia de investigación de siniestros.

Mentalmente, ¿cómo estructura el asegurado la decisión de defraudar?.

El proceso que lleva al asegurado a seguir un determinado tipo de comportamiento puede estructurarse mediante el árbol de decisión que se muestra en la Figura 1.

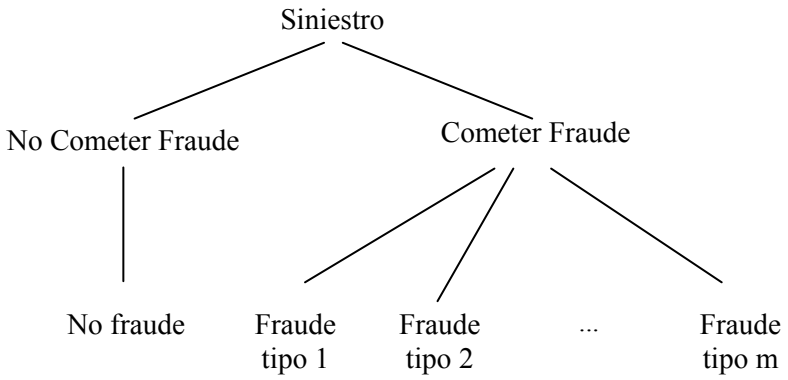


Figura 1

Atendiendo al esquema representado en la Figura 1, la elección de defraudar puede interpretarse como una elección realizada por etapas. De esta forma, el asegurado decide en primer lugar entre actuar honestamente o cometer fraude. Si opta por la primera alternativa, el siniestro es declarado de forma habitual y todo forma parte de la dinámica propia del sistema. Ahora bien, si opta por la segunda, un amplio abanico de posibilidades aparece a su alcance. Las circunstancias en las que se ha producido el siniestro, el tipo de cobertura, etc., pueden inducir al asegurado a realizar una elección concreta, en el sentido de cometer un determinado tipo de fraude.

Teniendo en cuenta este planteamiento, ¿cuál es nuestro objetivo?. Proporcionar a las compañías un instrumento que ayude a dirigir de forma adecuada la investigación de los siniestros. La aplicación de los modelos de elección múltiple permite determinar en primer lugar cuáles son los indicadores a estudiar para confirmar la sospecha de fraude en un expediente. En segundo lugar y, una vez se ha fundamentado dicha sospecha, el modelo indica cómo dirigir la investigación de cara a identificar un tipo de comportamiento fraudulento concreto, agilizándose el proceso de detección.

La aplicación de modelización logística multinomial y anidada permite estimar, de forma unificada, el árbol de decisión presentado en la Figura 1, sin considerar y considerando, respectivamente, la

existencia de un nivel intermedio de decisión. La elección finalmente realizada por el individuo estará fundamentada en un doble criterio: el deseo de maximizar la utilidad esperada de su comportamiento o la intención de minimizar los costes asociados al proceso de elección (tiempo empleado en la preparación del fraude,...). En esta elección intervendrán una serie de circunstancias analizadas dentro de nuestro modelo.

En Artís, Ayuso y Guillén (1999) se aplican, por primera vez, modelos logit anidados para cuantificar la probabilidad de aparición de diferentes tipos de fraude en una muestra de expedientes de siniestros del mercado asegurador español.

Consideremos, a continuación, la existencia de una secuencia jerárquica de decisiones y centrémonos en la especificación de un modelo logit anidado.

5.2.1 ESPECIFICACIÓN DE UN MODELO LOGIT ANIDADO.

La introducción del criterio de actuación del individuo en base a la maximización de la utilidad implicará considerar la premisa de que, entre varias alternativas, el asegurado siempre elegirá aquella que le reporte mayor utilidad. Según el modelo de utilidad aleatoria (McFadden, 1978), la utilidad que se deriva del comportamiento del individuo puede descomponerse en dos partes, una determinista (que genera la denominada *utilidad estricta*) y una aleatoria:

$$U_{i(cj)} = V_{i(cj)} + e_{i(cj)} .$$

La utilidad estricta, $V_{i(cj)}$, recoge una combinación lineal entre parámetros y variables explicativas propias del individuo y/o de la elección realizada (características del asegurado, del contrario, del siniestro, del vehículo,...). El término aleatorio o de error, $e_{i(cj)}$, recoge los efectos asociados a variables no consideradas en la parte determinista que pueden influir en la elección, así como las imperfecciones en la

percepción de la maximización de utilidad (McFadden, 1978; Maddala, 1983; Greene, 1997).

Teniendo en cuenta la existencia de una secuencia jerárquica de decisiones, la especificación de la función de utilidad implicará la inclusión de atributos propios de cada etapa de decisión, de forma que:

$$V_{i(cj)} = \beta' X_{i(cj)} + \alpha' Z_{i(c)}$$

siendo $X_{i(cj)}$ un vector de atributos específicos de cada elección final para el individuo *i*, $Z_{i(c)}$ el vector de variables observadas que varían sólo con la elección intermedia (decisión de defraudar o no defraudar, sin detallar el tipo de fraude que se realiza) y β y α los vectores de parámetros correspondientes.

La estimación de la probabilidad de elegir una determinada alternativa *cj*, será ahora el resultado de multiplicar dos probabilidades, una para cada nivel del árbol en el que nos situemos. Así,

$$P_{i(cj)} = P_{i(j|c)} P_{i(c)}$$

donde:

$P_{i(cj)}$ es la probabilidad de que el individuo *i* elija la alternativa *(cj)*,

$P_{i(c)}$ es la probabilidad de que *i* elija la alternativa intermedia *c*, y

$P_{i(j|c)}$ es la probabilidad condicionada de elegir la alternativa *j* una vez el individuo ya se ha decidido por la alternativa *c*.

Atendiendo a la especificación del modelo lógit multinomial, la probabilidad de que el individuo *i* elija una determinada opción final *j* (atendiendo a la elección previa de *c*) podrá definirse como:

$$P_{i(j|c)} = \frac{e^{V_{i(cj)}}}{\sum_{J=1}^{m_c} e^{V_{i(cJ)}}} = \frac{e^{\beta' X_{i(cj)}}}{\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}}},$$

donde *J* recoge el conjunto de elecciones posibles en la alternativa intermedia *c*. Por otro lado, la probabilidad asociada a la elección *c* atenderá a la expresión:

$$P_{i(c)} = \frac{\sum_{J=1}^{m_c} e^{V_{i(cJ)}}}{\sum_{b=1}^{C}\sum_{J'=1}^{m_b} e^{V_{i(bJ')}}} = \frac{e^{\alpha' Z_{i(c)} + (1-\sigma) I_{i(c)}}}{\sum_{b=1}^{C} e^{\alpha' Z_{i(b)} + (1-\sigma) I_{i(b)}}}$$

donde *b* recoge el conjunto de alternativas intermedias y *J'* el de alternativas dentro de cada opción intermedia. Además el subíndice *i* que indica el individuo especificado varía entre 1 y N. El valor, $I_{i(c)}$, denominado *valor inclusivo*, recoge la utilidad esperada agregada para un subconjunto de elección o conjunto de alternativas finales asociadas a la intermedia. Se define como,

$$I_{i(c)} = ln\left[\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}}\right].$$

La estimación de la probabilidad de elegir una determinada alternativa puede realizarse estimando los parámetros β a partir del modelo condicional ($P_{i(j|c)}$), determinado el valor inclusivo ($I_{i(c)}$) y estimando el vector α a partir de la probabilidad marginal $P_{i(c)}$. Esta aproximación no es sino una forma alternativa de modelizar la secuencia jerárquica de decisión, frente a lo que sería la estimación del modelo de forma completa (sin considerar la existencia de alternativas intermedias).

El coeficiente que acompaña al valor inclusivo recoge la posible correlación entre aquellas alternativas que, estando en el mismo nivel de decisión, forman parte de conjuntos de elección diferentes.

6. APLICACIÓN A UNA MUESTRA DE SINIESTROS EN EL RAMO DEL AUTOMÓVIL.

Veamos a continuación una aplicación de la modelización logística anidada, presentada en el apartado anterior, a una muestra de siniestros producidos en la cartera de automóviles de una entidad aseguradora española. Presentamos, en primer lugar, un análisis descriptivo de la información utilizada, para pasar, a continuación, a detallar los resultados obtenidos de estimar por máxima verosimilitud completa el árbol de decisión planteado (considerando de forma conjunta todo el proceso de elección).

6.1. LOS DATOS.

La muestra, relacionada por primera vez con expedientes en los que la existencia de fraude queda confirmada, recoge información para un total de 1995 siniestros, todos ellos acaecidos en España entre 1993 y 1996. La mitad de los expedientes han sido declarados legítimos por la entidad, mientras que en la otra mitad se ha demostrado la existencia de algún tipo de comportamiento fraudulento. Los tipos de fraude más frecuentes hacen referencia a dos situaciones alternativas: la falsa declaración del asegurado para eludir casos excluidos en la póliza o su falsa declaración para favorecer a un tercero. Otras situaciones están relacionadas con la falsa declaración del asegurado para obtener un beneficio sin intervención de un tercero, con la contratación de la póliza una vez ocurrido el accidente, con la ocultación de alcoholemia, con la existencia de fraude de taller, con la declaración de un falso conductor habitual para eludir los recargos o con la presentación de versiones cruzadas por parte de asegurado y contrario para cobrar ambos implicados.

Los expedientes de siniestros analizados se refieren a accidentes ocurridos en prácticamente todas las provincias españolas, recogiendo la oferta de coberturas ofrecidas por la compañía. No obstante y, para facilitar la interpretación, dichas coberturas han sido agrupadas en las categorías genéricas de “terceros”, “terceros más complementarios” y

"todo riesgo". Los diferentes tipos de vehículos quedan también suficientemente representados.

La modelización anidada desarrollada está relacionada con el escenario planteado en el árbol de decisión que aparece en la Figura 2.

MODELO. Fraude a favor del asegurado *versus* fraude a favor del contrario

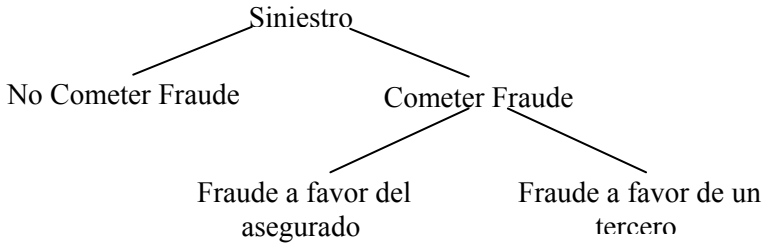


Figura 2

Como se desprende de la figura, en la modelización no se han considerado todos los posibles tipos de fraude que aparecen en la muestra y se ha optado por realizar una agrupación de los mismos en categorías más genéricas. El limitado número de expedientes englobados en algunas de las categorías iniciales y los problemas que de ello se derivarían en términos de no representatividad de las diferentes submuestras justifica tal actuación. Las posibilidades de agrupación son, no obstante, diversas, aunque este planteamiento es, a nuestro entender, el más general. El número total de expedientes analizados es, finalmente, de 1611 (998 casos legítimos, 299 casos con fraude a favor del asegurado y 314 con fraude a favor del contrario).

6.2. LAS VARIABLES Y SU ANÁLISIS DESCRIPTIVO.

Las variables explicativas introducidas en la especificación del modelo están relacionadas con aspectos diversos de la póliza, del siniestro y de las partes intervinientes en el mismo. Las diferencias

con los indicadores utilizados en el I.F.B. y en la Universidad de Montreal, como ya hemos avanzado en páginas anteriores, se dejen sentir fundamentalmente en dos aspectos importantes: por un lado, el hecho de que trabajemos con variables para las que la compañía dispone de información de manera inmediata; por otro, el que todas ellas gocen de carácter objetivo y no dependan de valoraciones hechas por los tramitadores.

Estos hechos diferenciadores poseen, lógicamente, ventajas y desventajas. Entre las ventajas se encuentra la rapidez con la que la compañía puede realizar la investigación y las consecuencias positivas que de ello se derivan teniendo en cuenta los cortos plazos de indemnización marcados por la aplicación de convenios (C.I.D.E.,...). Además, los tramitadores españoles no están acostumbrados a recoger, de forma rutinaria, el estado emocional del que declara, su elevado conocimiento de la jerga aseguradora o aspectos similares.

Sin embargo existen desventajas. La más importante es la dificultad para realizar un análisis detallado de la influencia que el fraude puede tener en la cobertura de daños corporales. Para ello resultaría necesario analizar variables que recogiesen el tratamiento médico aplicado, la duración del mismo, el tipo de daño (analizando, por ejemplo, la influencia de denominado "*efecto latigazo*[14]"). A diferencia de lo que ocurre en EE.UU., donde se da especial importancia a la detección de fraude en daños a la persona, en España la mayoría de fraudes detectados afectan a la cobertura de daños materiales. No obstante, nadie duda de la importancia que puede tener la aparición de comportamientos deshonestos en la cobertura por daños corporales que, aunque menos frecuentes, suelen llevar asociados indemnizaciones de mayor cuantía[15]. Además, se echa de menos una mayor cantidad de información relacionada con variables socio-económicas del asegurado (y/o del demandante) como pueden ser su situación laboral o el si forma parte o no de una familia.

[14] Con este término se recogen aquellos accidentes en los que se declaran daños menores, del tipo de dolores musculares, torceduras,...

[15] Estas conclusiones han sido ratificadas en los concursos de Detección de Fraudes en el Seguro que desde 1994 vienen celebrándose en España, organizados por I.C.E.A.

Las variables incluidas en la modelización y sus descriptivos quedan recogidos en las siguientes tablas:

Tabla 1
Variables en la base de datos

Nivel inferior del árbol de decisión	
RAMO	Cobertura a terceros (con o sin complementarios) igual a 1; 0, de otro modo
CULPA	Culpa del asegurado igual a 1; 0, en otro caso
ANTIGUO	Antigüedad del vehículo asegurado
EFECTO	Ocurrencia del siniestro cercana a la fecha de efecto de la póliza igual a 1; 0 en otro caso
Nivel superior del árbol de decisión	
HISTORIAL	Número de siniestros anteriores al declarado
FAMILIA	Existencia de parentesco entre los involucrados igual a 1; 0, en otro caso
POLICÍA	Intervención de policía en el accidente igual a 1; 0, en otro caso
RELATO	Descripción sospechosa de los hechos igual a 1; 0, en caso contrario

Tabla 2
Estadísticos descriptivos

Variable	Total muestra[a]		Fraude		No Fraude	
	Media	Desv. Std	Media	Desv. Std	Media	Desv. Estandar
RAMO	0.90	0.30	0.94	0.24	0.87	0.34
CULPA	0.30	0.46	0.52	0.50	0.17	0.38
ANTIGUO	6.17	0.45	6.31	4.71	6.09	4.42
EFECTO	0.01	0.09	0.01	0.11	0.01	0.08
HISTORIAL	1.40	1.81	1.69	2.02	1.22	1.65
FAMILIA	0.06	0.23	0.10	0.30	0.03	0.17
POLICÍA	0.13	0.34	0.04	0.19	0.19	0.39
RELATO	0.58	0.49	0.63	0.48	0.55	0.50

[a]Estadísticos muestrales no ponderados

6.3. RESULTADOS DE LA ESTIMACIÓN.

Los resultados obtenidos en la estimación del modelo propuesto (Figura 2) para el mercado español se muestran en la tabla 3. Éstos confirman la existencia de variables con diferente influencia en la explicación de los diversos tipos de fraude. El modelo se ha estimado por el método de Máxima Verosimilitud Completa y se ha tenido en cuenta el efecto de la sobre-representación de los siniestros fraudulentos.

Las variables consideradas no cambian con la alternativa elegida. La variable dependiente es el tipo de siniestro atendiendo a su clasificación en legítimo, fraude a favor del asegurado y fraude a favor del contrario.

Tabla 3

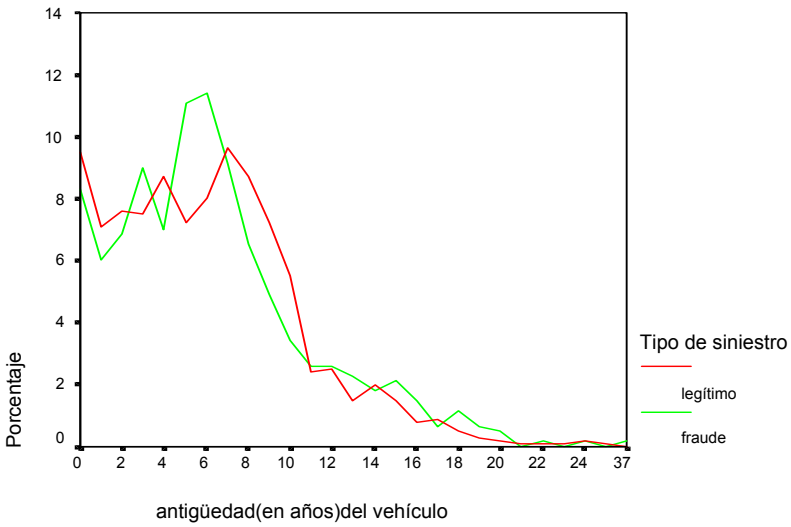
Estimación del modelo logit anidado con ponderaciones

Variable	Coeficiente	t-test	P-value
Nivel inferior (fraude)			
Constante	1.573	4.606	0.000
RAMO	0.689	2.705	0.007
CULPA	2.006	8.008	0.000
ANTIGUO	0.070	2.039	0.041
ANTIGUO (al cuadrado)	-0.004	-1.882	0.060
EFECTO	0.867	1.697	0.090
Nivel superior:			
Constante	-5.990	-8.540	0.000
HISTORIAL	0.155	4.385	0.000
FAMILIA	1.414	5.073	0.000
POLICÍA	-1.624	-5.620	0.000
RELATO	0.509	3.686	0.000
Valores Inclusivos			
NO FRAUDE	1.000	----	----
FRAUDE	1.835	5.614	0.000
Número de observaciones[a]	1611	Chi-cuadrado	1264.49
Función de log verosimilitud	-909.31	Grados de libertad	12
Log verosimilitud restringida	-1541.56	Nivel Significación	0.00

[a] Cada individuo es replicado una vez por cada elección.

Los coeficientes estimados son significativos al 5% (con excepción de los parámetros que acompañan a las variables *antiguo (al cuadrado)* y *efecto* que lo son al 10%) y tienen los signos esperados.

Los resultados permiten concluir que cuanto más antiguo sea el vehículo asegurado mayor es la probabilidad de que el asegurado actúe en beneficio propio y no en beneficio del conductor contrario, aunque se ha detectado la existencia de un punto de inflexión a partir del cual dicha probabilidad disminuye, como consecuencia probablemente del bajo valor venal del vehículo. Este resultado se pone de manifiesto con el análisis gráfico de la variable en el que se observa claramente el punto de inflexión mencionado (Gráfico 1). Asimismo, cuando el asegurado declara el siniestro en fecha cercana a la contratación de la póliza, aumenta la probabilidad de que defraude.



Cuando el asegurado tiene contratada únicamente cobertura a terceros y cuando admite la culpa del siniestro aumenta la probabilidad de que defraude en beneficio del contrario, como es de esperar. Estos dos resultados, al tratarse del nivel inferior del árbol (Figura 2) suponen que la decisión sobre defraudar es clara.

En el nivel superior, los factores especificados influyen significativamente en la aparición de fraude. A modo de ejemplo, el hecho de que no sea el primer siniestro declarado por el asegurado, el parentesco entre los implicados o bien la existencia de una descripción sospechosa de los hechos ocurridos (relatos relacionados con la realización de determinadas maniobras, como marcha atrás, aparcamiento,...), son indicadores relevantes de la posible presencia de comportamientos fraudulentos. En caso de que en la declaración del siniestro conste que intervino la autoridad, la probabilidad de que el siniestro contenga fraude disminuye significativamente.

La calidad del ajuste, medida a partir de la siguiente tabla de frecuencias de clasificación, muestra resultados satisfactorios,

Tabla 4
Frecuencias de clasificación (punto de corte=0.15)

	Elección predicha			
	Legítimo	Beneficio propio	Beneficio tercero	Total
Elección Observada				
Legítimo	796	59	143	998
Beneficio propio	172	53	74	299
Beneficio tercero	62	16	236	314
Total	1030	128	453	1611

Como resultado del proceso obtenemos la probabilidad estimada o ajustada de que en el expediente exista cada tipo de comportamiento. El siniestro es clasificado en aquella categoría para la que la probabilidad estimada es máxima. La comparación entre la clasificación real u observada para el expediente y la derivada de la modelización permite obtener una primera conclusión de la capacidad que el modelo posee para reproducir el comportamiento observado en

la muestra. En el modelo especificado, el porcentaje de casos correctamente clasificados es del 67.3%.

Los resultados obtenidos ponen de manifiesto que de cada 100 expedientes sujetos a investigación, aproximadamente 67 serán correctamente clasificados en su categoría correspondiente. Para los 33 restantes el error cometido podrá ser de tres tipos:

1. Siniestros legítimos incorrectamente clasificados como fraudulentos (de uno u otro tipo).
2. Siniestros fraudulentos de un tipo incorrectamente clasificados en el otro.
3. Siniestros fraudulentos de uno u otro tipo incorrectamente clasificados como legítimos.

¿Qué consecuencias traerán cada uno de estos errores?.

En el primero de los casos la compañía abrirá un proceso de investigación sobre un conjunto de expedientes sobre los que será incapaz de demostrar la existencia de fraude dado que éste no existe. Los principales inconvenientes estarán asociados a los costes derivados de la investigación y al tiempo empleado en la misma. En cualquiera de los casos, el tramitador nunca debe confirmar la existencia de fraude hasta que no disponga de todas las pruebas necesarias para hacerlo.

En el segundo de los casos, las consecuencias del error cometido serán menos importantes: la compañía acabará detectando comportamiento fraudulento.

Es en el tercero y último de los casos planteados donde posiblemente se produzca el error más importante. Como consecuencia del mismo, un determinado número de expedientes fraudulentos acabarán siendo clasificados por el modelo como no fraudulentos. Lógicamente, sobre los mismos la compañía no realizará ningún tipo de investigación y acabará pagando la indemnización como si de siniestros legítimos se tratara.

La optimización del punto de corte (probabilidad predicha a partir de la cual el siniestro será considerado fraudulento) puede ayudar a mejorar el porcentaje de casos fraudulentos correctamente clasificados.

Todo el proceso estará influenciado por los costes asociados a la investigación. En el trabajo realizado se supone que la optimalidad de un sistema de clasificación viene dada por su capacidad para detectar los fraudes pero podría ser mejorado si se dispusiera de información sobre los diferentes costes que genera la investigación del siniestro. Los resultados presentados por I.C.E.A (2000) indican que las cantidades invertidas en la investigación de los siniestros son mínimas, teniendo en cuenta las elevadas cantidades finalmente ahorradas por las entidades[16]. Desde este punto de vista y, siempre con carácter preliminar, podría señalarse que los costes iniciales y de mantenimiento derivados de la implementación de un mecanismo de detección se verán, con gran probabilidad, sobrepasados por los beneficios inducidos por éste.

Las conclusiones extraídas sobre la calidad del ajuste del modelo logit anidado permiten caracterizarlo como técnica de predicción aceptable. Además queda también demostrada su capacidad como herramienta explicativa. Es de destacar el hecho de que gracias a su aplicación seamos capaces de proponer un modelo para el proceso de decisión que lleva al asegurado a cometer o no fraude. Además determinamos qué variables influyen en cada elección, siendo destacable la intervención de las mismas en las diferentes etapas de decisión. Así, la entidad puede saber *a priori* qué variables debe investigar, no sólo para detectar fraude, genéricamente hablando, sino también para

[16] Según los resultados obtenidos del VI Concurso de Detección de Fraudes organizado por I.C.E.A. en 1999 (I.C.E.A., 2000), las 25 entidades aseguradoras participantes descubrieron un total de 24473 intentos de fraude (el seguro del automóvil concentra aproximadamente el 93%). El fraude bruto evitado ascendió a 6178 millones de pesetas (el 80% en automóviles) mientras que las acciones de investigación y prevención supusieron únicamente un gasto aproximado del 2.53% de las reclamaciones (aproximadamente 182 millones de pesetas). En autos (22703 casos) la cuantía de fraude detectado supuso un ahorro de 4941 millones de pesetas, siendo la inversión media en cada caso investigado de 5499 pesetas aproximadamente.

detectar aquellos comportamientos que con mayor frecuencia aparecen en la cartera de asegurados. La obtención de muestras más representativas para los tipos de fraude analizados ayudaría, posiblemente, a aumentar la calidad predictiva del modelo planteado.

El análisis coste beneficio asociado a la implementación del sistema propuesto y el diseño de un mecanismo de detección que tenga en cuenta la vida del siniestro (determinación de los indicadores de fraude que inciden en cada etapa de la tramitación) son dos de las líneas futuras de investigación propuestas por el equipo de trabajo. No obstante, la obtención de resultados para ambos objetivos vendrá condicionada por dos hechos fundamentales: la disponibilidad de información exhaustiva para los costes asociados al siniestro y la posibilidad de trabajar con muestras de expedientes de siniestros abiertos (consideración de todo el proceso de tramitación, desde la apertura del expediente hasta su liquidación).

CONCLUSIONES

La aplicación de métodos cuantitativos dirigidos a detectar y controlar el fraude queda más que justificada. La determinación de las variables sobre las que la compañía debe dirigir la investigación de cara a detectar un determinado tipo de comportamiento y, la implementación de un mecanismo de detección que cuantifique la probabilidad de que un expediente contenga fraude, son dos de los pasos sugeridos para realizar una correcta política de control. La aplicación de métodos estadísticos y econométricos puede constituirse en una herramienta básica, junto a la necesaria motivación del personal de la entidad, en la detección de casos fraudulentos.

La participación de la comunidad científica en el tratamiento del problema cada día es más notable. Las diferencias de nuestro trabajo con las aportaciones realizadas por otros autores se hacen patentes tanto en el tratamiento metodológico aplicado y en los indicadores de fraude utilizados como en el hecho de considerar un conjunto de alternativas de elección más amplio que la dicotomía fraude/no fraude.

El Insurance Fraud Bureau de Massachusetts, de la mano de Richard Derrig, utiliza el modelo de regresión lineal múltiple para determinar que indicadores deben investigarse como señales de fraude. Su principal objetivo es señalar las características básicas que permiten fundamentar la existencia de sospecha de comportamiento fraudulento. En la clasificación del siniestro se diferencian cuatro situaciones alternativas: siniestro legítimo, siniestro con fraude planeado (siniestro provocado); siniestro con fraude oportunista (siniestro real pero sin existencia de daños) y siniestro con hinchamiento de los gastos médicos derivados (el siniestro es real, existen daños a la persona pero el asegurado los exagera injustificadamente).

La Universidad de Montreal presenta en sus estudios una mayor similitud con los realizados por nuestro equipo de trabajo. La cuantificación de la probabilidad de existencia o sospecha de fraude (previa selección de los indicadores a introducir en la especificación del modelo y el análisis estadístico de los mismos) y la determinación del criterio probabilístico óptimo a utilizar por la compañía para clasificar correctamente los siniestros en su categoría respectiva son los principales resultados del análisis realizado por Belhadji y Dionne en 1997. Sin embargo las diferencias con nuestro trabajo se hacen patentes en la vertiente metodológica aplicada. En sus estudios todos los resultados aparecen asociados a la dicotomía fraude/no fraude sin que consideren la existencia de diferentes tipos de comportamientos fraudulentos ni tengan en cuenta la posibilidad de modelizar la elección de defraudar como una elección realizada por etapas. La técnica econométrica utilizada es, de esta forma, más sencilla, proponiendo el uso de un modelo próbit. La definición de patrones de comportamiento queda limitada al caso general de presencia-ausencia de fraude sin que sea posible determinar la existencia de atributos directamente relacionados con sus diferentes formas de manifestarse.

Nuestra investigación apuesta por el uso de modelos de elección probabilística, logísticos multinomiales y anidados, para cuantificar la probabilidad de que un expediente contenga un determinado tipo de fraude. Además de trabajar con variables objetivas y con información inmediatamente disponible para la compañía, trabajamos con una

muestra de expedientes de siniestros que contienen fraude detectado (los otros grupos de trabajo utilizan muestras en las que existe sospecha de fraude). El diseño de un método dirigido a optimizar el criterio probabilístico utilizado para clasificar un siniestro en una determinada categoría, en modelos con tres alternativas finales de elección, permite mejorar notablemente la calidad predictiva de los mismos. La modelización de la decisión de defraudar como una decisión realizada por etapas permite determinar, en primer lugar, qué variables han de ser investigadas para fundamentar la sospecha de fraude y, en segundo lugar, cuáles confirman la existencia de un determinado tipo de comportamiento fraudulento. En definitiva, proporciona una herramienta para dirigir de forma adecuada la investigación de los siniestros.

BIBLIOGRAFÍA

[1]- A.I.B. (1997) 1997 **Annual Report**. Automobile Insurers Bureau of Massachusetts. Boston.

[2]- ARTÍS, M., M. AYUSO y M. GUILLÉN (1999) **Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market.** Insurance: Mathematics and Economics, 24, 67-81.

[3]- AYUSO, M. y M. GUILLÉN (1995a) **El Fraude en el Seguro del Automóvil.** Trabajo de Investigación realizado dentro del Programa de Doctorado Economía y Territorio: Análisis Cuantitativo. Departamento de Econometría, Estadística y Economía Española. Universidad de Barcelona.

[4]- AYUSO, M. y M. GUILLÉN (1995b) **Modelos de Elección del Fraude en el Seguro de Automóviles**. Actas de la IX Reunión de la Asociación Científica Europea de Economía Aplicada ASEPELT-España celebrada en Santiago de Compostela, 22 y 23 de junio de 1995, vol. II (Economía Sectorial), 383-394.

[5]-AYUSO, M. y M. GUILLÉN (1999) **Modelos de Detección de Fraude en el Seguro del Automóvil**. Cuadernos Actuariales, vol. 8, 135-150.

[6]- BELHADJI, E.B. Y G. DIONNE (1997) **Development of an Expert System for the Automatic Detection of Automobile Insurance**

Fraud. Working Paper 97-06. École des Hautes Études Commerciales. Université de Montréal.

[7]- BROCKETT, P.L., X. XIA y R. DERRIG (1995) **Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud**. Journal of Risk and Insurance, vol. 65, nº 2, 245-274.

[8]- C.E.S. (1992) **El Fraude en el Seguro de Automóviles**. Centro de Estudios del Seguro. Madrid.

[9]- COBO, P. (1993) **Manual de Investigación de Siniestros y Lucha contra el Fraude en el Seguro de Automóviles**. Ed. Mapfre, Madrid.

[10]- COMITÉ EUROPEO DE SEGUROS (1996) **Le Guide de l'anti-fraude à l'assurance en Europe**. CEA Info, Hors - Série nº 4, Mai 1996.

[11]- CUMMINS, J.D. y S. TENNYSON (1996) **Moral Hazard in Insurance Claiming: Evidence from Automobile Insurance**. Journal of Risk and Uncertainty, vol. 12, nº 1, 29-50.

[12]- DERRIG, R.A. y K.M. OSTASZEWSKI (1995), **Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification**, Journal of Risk and Insurance, vol. 62, nº 3, 447-482.

[13]- DERRIG, R.A. Y H.I. WEISBERG (1998) **AIB PIP Claim Screening Experiment Final Report. Understanding and Improving the Claim Investigation Process**, DOI Docket R98-41, Boston.

[14]- I.C.E.A. (2000) **El Fraude al Seguro Español. Acciones para combatirlo**. Asociación I.C.E.A., Informe nº 797, junio 2000.

[15]- WEISBERG, H.I. Y R.A. DERRIG (1991) **Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Liability Claims in Massachusetts**. Journal of Insurance Regulation, vol. 9, nº 4, 497-541.

[16]- WEISBERG, H.I. Y R.A. DERRIG (1993) **Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims**. AIB Cost Containment/Fraud Filing, 49-82.