uc3m | Universidad Carlos III de Madrid

Máster Universitario en Ciencias Actuariales y Financieras

2018/2019

Trabajo Fin de Máster

GEORREFERENCIACIÓN DE LA TASA DE ROBO EN SEGUROS DE HOGAR MEDIANTE LOS APLICATIVOS DE PYTHON Y CARTO

Mª Elena Fernández Boza

Tutor/es

José Miguel Rodríguez-Pardo Del Castillo

Jesús Ramón Simón Del Potro

Madrid, 2019

RESUMEN

En este trabajo de fin de máster se trata de explicar las posibles causas en las variaciones de la tasa de robos en viviendas, siendo una de las partes fundamentales para el cálculo de la tarifa en seguros de hogar, en concreto, para la cobertura de robo.

En el presente documento se utilizan técnicas de modelización estadística-actuarial, como lo son los modelos clásicos *GLM* (*Generalized Linear Models*) asociados a técnicas de los *Sistemas de Información Geográfica* (*SIG*) que permiten localizar las zonas analizadas que presenta un mayor o menor índice de criminalidad para este tipo de delitos en función de las variables añadidas a las modelos.

En resumen, esta Tesis contiene una combinación de las técnicas tradicionales con las nuevas tecnologías aplicables al sector asegurador, sirviendo de base en futuras investigaciones.

PALABRAS CLAVE: Modelos Lineales Generalizados, Seguro de hogar, Sistemas de Información Geográfica, SIG, Tarificación, Poisson, Binomial Negativa.

ABSTRACT

This master´s Thesis tries to explain the possible causes for the variations of the robbery rate in house insurance used to the pricing, mainly, to stealing covering.

In the present document have been used techniques of statistics-actuarial modelization such as the classic models *GLM* (*Generalized Linear Models*) joined to *SIG* (*Geographical Information Systems*) which allow alocate the analysed areas that have a higher or lower criminality index for this kins of offenses based on the variable added to the models

In short, this master´s Thesis presents a combination of traditional techniques joined to the new technologies applied to the insurance sector, being a base of futures researchs.

KEYWORDS: Generalized Linear Models, Home Insurances, Geographic Information System, GIS, Pricing, Poisson, Negative Binomial.

AGRADECIMIENTOS

Quiero hacer constar que la finalización de esta etapa no hubiese sido posible sin el apoyo incondicional de todas las personas que me rodean. ¡Gracias!

A mis tutores D. José Miguel Rodríguez Pardo y D. Jesús Ramón Simón del Potro, quienes me han orientado en todo momento ante las dificultades que se han ido presentando a lo largo de la realización del presente trabajo, sin importarles ni la hora, ni el día de la semana.

A mis compañeros de Deloitte S.L. del departamento de *Risk Advisory*, que a pesar de llevar poco tiempo me han demostrado ser como una gran familia.

A dos personas anónimas a quiénes les tengo un especial respeto y admiración; por la cercanía ofrecida, la confianza depositada y por haberme demostrado que por medio de las anotaciones es como mayormente se aprende, siendo el motor de ese aprendizaje los consejos basados en sus experiencias vividas como grandes personas y profesionales, a su vez.

A mi familia, abuelos/as, tios/as, primos/as, suegro/a…, especialmente a mi pareja y a mi abuelo Kiko, siendo los que me han transmitido las fuerzas en momentos de debilidad.

Los mayores partícipes mis padres, junto con mi hermano Hugo, quienes me han demostrado que la distancia no es un obstáculo comparado con el apoyo, los valores y los consejos que me han transmitido para alcanzar mis metas, haciéndome ver que nada en imposible con esfuerzo y sacrificio si te lo propones.

INDICE

INDICE DE TABLAS

INDICE DE GRÁFICOS

INDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

1.1. PRINCIPALES PREOCUPACIONES DEL SECTOR ASEGURADOR

El sector asegurador es considerado como un sector bastante conservador. El dinamismo del entorno conlleva a que las preferencias de los clientes cambien a un ritmo exponencial, sobre todo, debido a la importancia en la tecnología para la nueva generación "millenials".

Por un lado, el tradicional modelo de seguro debe de estar enfocado a canalizar de forma eficaz los productos ofertados, sin obviar la diversidad de perfiles que se presentan en el mercado, pero a su vez, haciendo hincapié en maximizar la experiencia de cada perfil de asegurador.

Por otro lado, la diferenciación respecto a los competidores es un factor imprescindible para garantizar la fidelización por parte de los clientes. Sin embargo, los cambios en el desarrollo de las nuevas tendencias suponen para las compañías un nuevo reto y altas necesidades de adaptarse al mercado de la economía digital, siendo las grandes multinacionales las primeras en emergerse en la transformación y, por tanto, marcando la diferencia en las técnicas aplicadas en el sector. Para determinar el impacto que supone la introducción de nuevas metodologías, hay que tener en cuenta las principales características que presenta la actividad del sector seguros:

- Grandes volúmenes de datos.
- Alto marco normativo.
- Múltiples canales de actuación: distribución directa, mediadores, banca-seguros, etc.
- Procesos altamente relacionados con el factor humano.

A todo ello, hay que sumarle la gran inversión que conlleva implementar nuevas técnicas a futuro. Esto hace que las compañías no tengan claro el nuevo enfoque estratégico y, no sean susceptibles todavía a introducir procesos de automatización o innovación tecnológica en sus diferentes fases en la estructura de su actividad empresarial.

1.2. APLICACIONES DE LAS NUEVAS TECNOLOGÍAS. SOLUCIÓN AL CAMBIO

La revolución digital es un aspecto que está presente en la actualidad siendo de vital importancia para el sector asegurador. Que una empresa sea eficiente en sus procesos conlleva a obtener una ventaja competitiva y, en consecuencia, una optimización de las funciones que permite minimizar riesgos operacionales, disminuir los costes y aumentar la experiencia recibida por parte del cliente.

La innovación y automatización tecnológica se ha convertido en un pilar fundamental para alcanzar el éxito en las compañías.

A continuación, se muestra un esquema que refleja el impacto monetario de aplicar procesos de automatización (RPA[1]) en un horizonte temporal a corto plazo.

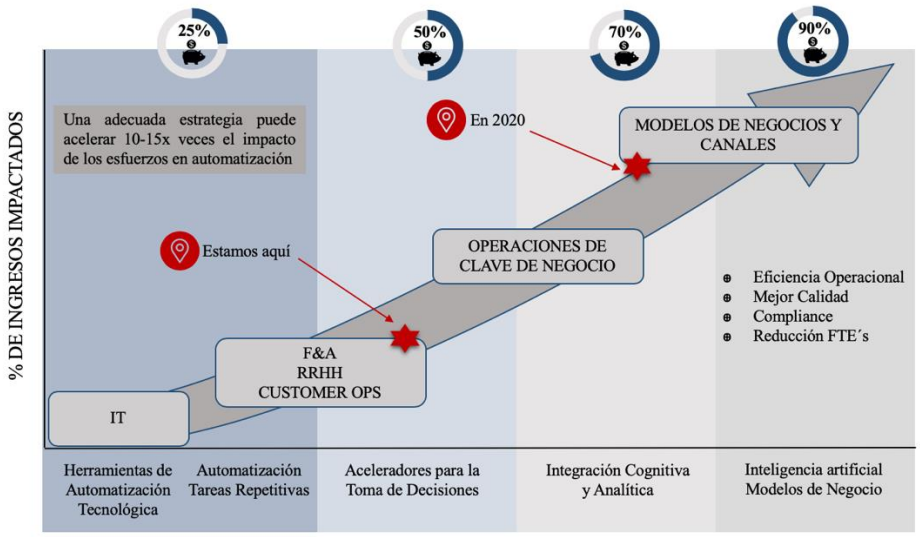


Ilustración 1. Desarrollo en Automatización. Elaboración Propia inspirada por (Avasant, Mayo 2017)

Un ejemplo revolucionario de estos tipos de tecnologías son los novedosos "*Chatbots*" creados con técnicas de inteligencia artificial que permiten que el cliente interactúe con un robot capaz de realizar cualquier tipo de tarea como, por ejemplo, proporcionarle información para la contratación de un seguro.

Por todo ello, hay que destacar que en un periodo corto de tiempo la revolución del tratamiento de grandes volúmenes de datos mediante la implementación de técnicas de machine learning, deep learning y, los sistemas de información geográfica (*SIG*), entre otras, permitirá a las compañías tomar decisiones de forma eficaz y eficiente pudiendo

[1] Robotic Process Automation.

automatizar los diferentes procesos vinculados a la toma de decisiones de la actividad aseguradora y, en consecuencia, incrementar sus beneficios.

El *Gráfico 1* muestra las principales fases del proceso de las compañías aseguradoras donde están utilizando RPA actualmente y la posibilidad de incorporación futura en las mismas.

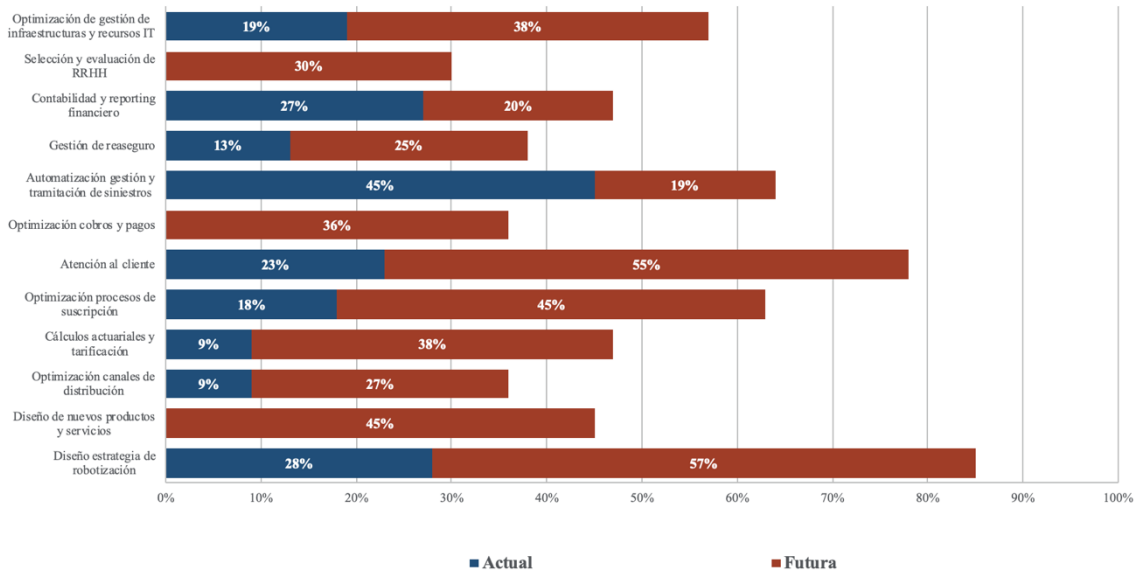


Gráfico 1. Iniciativas RPA actuales y futuras. Elaboración Propia inspirada por (Avasant, Mayo 2017)

Las tareas de atención al cliente, procesos de suscripción y el diseño de nuevos productos y servicios son las fases con mayor proporción a iniciarse en el cambio futuro de introducir procesos de automatización, seguido del cálculo en la tarificación de las pólizas, siendo fases muy ligadas al trato relacionado con clientes.

Otro aspecto relevante en el cambio tecnológico es el valor añadido que proporciona trabajar con *sistemas de información geográfica (SIG)*, ya que son técnicas que permiten utilizar información espacial pudiendo ser incorporadas como variables para la fase de tarificación de algunos tipos de seguros, como son hogar y autos.

1.3. OBJETIVOS A DESARROLLAR

La capacidad analítica forma parte de la rutina de cualquier actuario, siendo una de las tareas fundamentales la modelización de los riesgos asociados a la fase de tarificación de primas para cuantificarlas de forma adecuada, entre otras.

Con el presente trabajo se pretende conseguir un modelo que proyecte de forma adecuada la tasa de robo en viviendas derivado del seguro de hogar obteniendo aquellas variables

que aportan mayor información a la variable respuesta y, en consecuencia, comprender sus causas.

Además, se utiliza *Sistemas de Información Geográfica* (*SIG*) que permiten evaluar aquellas zonas con mayor potencial de criminalidad en este tipo de delitos, tomando como capa los distritos para la ciudad de Madrid.

La propuesta de esta temática deriva de la importancia de comprender e implementar nuevas técnicas que permitan un análisis más profundo de los modelos actuariales en base a la explotación de los datos. La utilización geoespacial permite que los modelos se enriquezcan aportando una mayor precisión en los resultados obtenidos y, en consecuencia, un mayor beneficio para la compañía.

2. EL SECTOR ASEGURADOR EN ESPAÑA

2.1. VISIÓN GENERAL DEL SECTOR EN CIFRAS

El sector asegurador ocupa una posición muy destacada en la economía de España. El control de las funciones de la actividad aseguradora se regula por la *Dirección General de Seguros y Fondos de Pensiones (DGSFP)* que se encarga de que las entidades que se dedican a dicho sector cumplan con las regulaciones legales derivadas de la propia actividad.

Para que se materialice un contrato de seguro es necesario que exista:

- ◊ *Compañía de seguros* legalmente establecida, que adopta la figura de asegurador (reflejado en *LOSSEAR*).
- ◊ *Riesgo* asumible por el asegurador, transferido por el asegurado a cambio de una prima.
- ◊ *Prima*. Definida como la cuantificación del riesgo o precio del seguro.

De esta forma, (Art. 1, Ley 50/1980, 8 de octubre, de Contrato de Seguro, 1980) establece que:

"el contrato de seguro es aquel por el que el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura a indemnizar, dentro de los límites pactados, el daño producido al asegurado o a satisfacer un capital, una renta u otras prestaciones convenidas".

A continuación, se muestra a modo resumen la tendencia cuantitativa que presenta el sector con datos extraídos del último informe publicado por la DGSFP pertenecientes al año 2017. Cabe destacar que la aportación más significativa en el sector es el valor proporcionado por las primas devengadas brutas (PDB)[2].

2 Según especifica (DGSFP, 2017) "Primas devengadas brutas (PDB) = Primas devengadas de seguro directo + Primas devengadas de reaseguro aceptado".

Tabla 1.El seguro en la economía española[3]

	2013	2014	2015	2016	2017
PDB	56.263	56.016	57.073	64.920	64.514
PIB a p.m.	1.025.634	1.037.025	1.075.639	1.118.522	1.163.662
PDB/PIB	5,49%	5,40%	5,31%	5,80%	5,54%
PDB/habitantes	1.194	1.198	1.224	1.394	1.385

Fuente: (DGSFP, 2017)

Se aprecia que el volumen de primas presenta una clara tendencia de crecimiento. Sin embargo, en el último año se observa un breve descenso incentivado por un aumento en el valor del PIB y, un incremento en el valor obtenido por las primas devengadas brutas en dicho periodo.

Por otro lado, si analizamos el sector asegurador en función de los diferentes ramos, los resultados obtenidos son los siguientes:

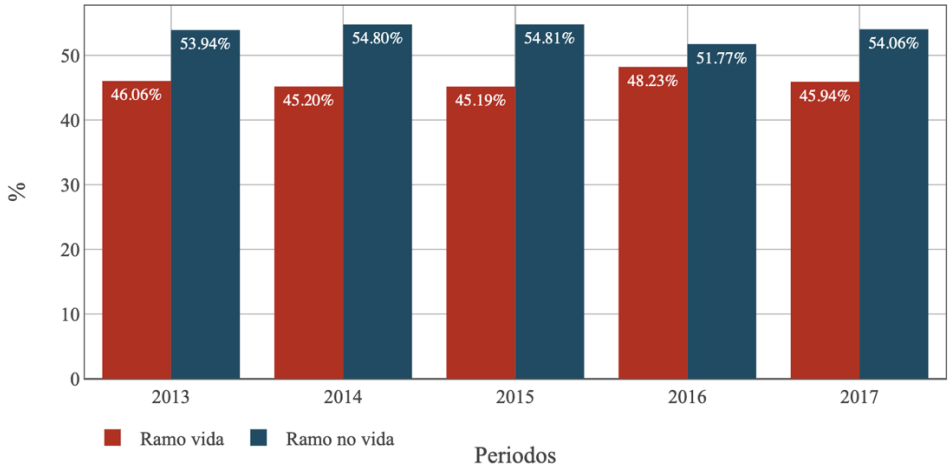


Gráfico 2. Comparativa Ramos. Elaboración Propia (DGSFP, 2017)

El ramo de no vida tiene mayor impacto que el ramo de vida respecto al total del sector, manteniendo valores constantes a lo largo del periodo a excepción del año 2016, donde se aprecia una leve disminución y, en contraposición un aumento del ramo de vida respecto a periodos anteriores. En el último periodo, se observa como ambos ramos adoptan de nuevo valores muy similares a la etapa comprendida entre 2013-2015, es decir, el valor del ramo de vida presenta un comportamiento decreciente, mientras que el valor del ramo no vida creciente.

3 Datos expresados en millones de €.

Tabla 2. Primas Devengadas Brutas y Variación Total[4]

	2013	2014	2015	2016	2017
PDB	56.263	56.016	57.073	64.920	64.514
Ramo vida	25.913	25.321	25.791	31.309	29.639
Ramo no vida	30.350	30.695	31.282	33.612	34.875
PDB a p.m.	1.025.634	1.037.025	1.075.639	1.118.522	1.163.662
Variación total	-2,57%	-0,44%	1,89%	13,75%	-0,63%
Variación vida	-2,98%	-2,29%	1,86%	21,39%	-5,33%
Var. no vida	-2,22%	1,14%	1,91%	7,45%	3,76%
Var. PIB a p.m.	-1,36%	1,11%	3,72%	3,99%	4,04%

Fuente: (DGSFP, 2017)

Se puede observar que el conjunto de variaciones a partir del periodo 2013 comienzan a experimentar una transformación, es decir, los valores negativos se transforman a menores disminuciones en el valor hasta convertirse en valores positivos en el año 2016. En el último periodo equivalente al año 2017 la variación que pertenece al ramo total y al ramo vida decrece, mientras que el ramo de vida presenta un incremento, aunque con un valor inferior a la variación del PIB.

2.2. LA IMPORTANCIA CUANTITATIVA DE LOS DIFERENTES RAMOS EN LA ACTUALIDAD

El sector asegurador divide según la normativa su actividad en diferentes ramos, entendidos éstos, como un conjunto de seguros que cubren determinados riesgos de naturalezas semejantes. Además, cada ramo está compuesto por seguros con modalidades o categorías diferentes, las cuales se diferencian por cubrir riesgos específicos. De esta forma, la normativa los clasifica en:

- ***Seguros de vida***. "Por el seguro de vida el asegurador se obliga, mediante el cobro de la prima estipulada y dentro de los límites establecidos en la Ley y en el contrato, a satisfacer al beneficiario un capital, una renta u otras prestaciones convenidas, en el caso de muerte o bien de supervivencia del asegurado, o de ambos eventos conjuntamente", definición establecida según (Art. 83, Ley 50/1980, 8 de octubre, de Contrato de Seguro, 1980)

[4] Datos expresados en millones de €. Variación expresada en %.

- ***Seguros de no vida.*** El fin de estos seguros es reparar las pérdidas ocasionadas tanto en el patrimonio del tomador[5] como las pérdidas materiales directamente sufridas por el acaecimiento de un siniestro. Son los denominados seguros sobre cosas y de responsabilidad.

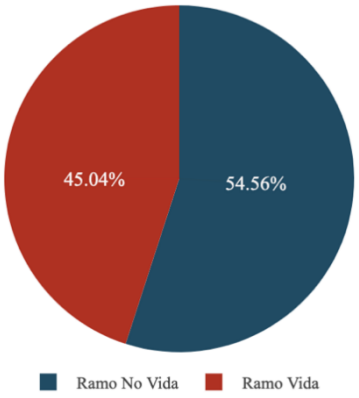


Gráfico 3. Volumen Estimado de Primas por Ramo. Elaboración Propia (ICEA, 2019)

El ramo de no vida ocupa una mayor proporción en el mercado que el ramo de vida, aunque la diferencia entre ambos no es muy dispar. Esto es debido a que las modalidades que integran dicho ramo tienen mayores índices de contratación a causa de que son seguros que cumplen un principio indemnizatorio. Si los resultados obtenidos los comparamos con el *Gráfico 2*, se aprecia como no existe variación de los sectores de un periodo a otro.

2.2.1. CLASIFICACIÓN DE LOS DIFERENTES TIPOS DE SEGUROS

El acaecimiento de un siniestro se considera un evento aleatorio, por esta razón, la clasificación de los diferentes tipos de seguros nace bajo el supuesto de la existencia de la diversidad de riesgos específicos que se presentan. Cada tipo de seguro presenta unas garantías específicas y, a su vez, diferentes en función del riesgo que se pretenda cubrir, pero todos tienen en común que permiten adquirir una promesa futura.

La *Ilustración 2* muestra las principales tipologías de seguros que se ofertan en el mercado español.

[5] Persona que contrata el seguro, es decir, firmante de la póliza. Nótese que la figura del tomador y asegurado puede ser diferente.

Ilustración 2. Tipología de Seguros. Fuente: Elaboración propia (Fundación MAPFRE, 2019)

Los ***seguros contra daños patrimoniales*** cumplen un principio estrictamente indemnizatorio, es decir, en caso de que se materialice un siniestro y, en consecuencia, el daño, ese daño es económicamente indemnizable siendo el importe máximo a percibir la suma asegurada (Art. 27, Ley 50/1980, 8 de octubre, de Contrato de Seguro, 1980).

En los ***seguros de personas***, más que percibir una indemnización, se pretende compensar la pérdida económica o la necesidad de la contingencia que se puede establecer (enfermedad, dependencia, jubilación, incapacidad parcial o total o fallecimiento). La principal característica en estos seguros es que existe una necesidad presunta y, por tanto, la suma asegurada es preestablecida a priori cumpliendo una función más que indemnizatoria, de previsión o social.

Teniendo en cuenta la clasificación anterior, es importante conocer la proporción equivalente a cada uno de las modalidades dentro de los diferentes ramos ya que permite conocer las preferencias reales de los asegurados en función de las demandas del mercado. En este caso las modalidades analizadas son[6]:

[6] Nótese que para mayor información de las subcategorías de cada modalidad se puede consultar el Anexo A

Gráfico 2. Comparativa Ramos. Elaboración Propia

.

◊ Ramo vida: ahorro y riesgo.

◊ Ramo no vida: automóviles, multirriesgos, salud y otros.

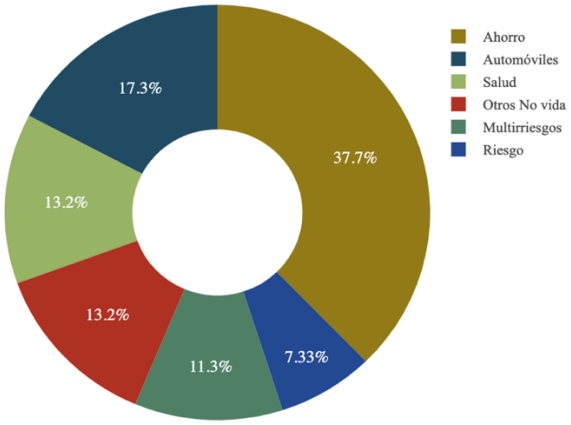


Gráfico 4. Volumen Estimado Primas por Modalidad. Año 2018. Elaboración Propia (ICEA, 2019)

La modalidad de ahorro es la que presenta un mayor porcentaje, siendo un dato que manifiesta la preocupación de la población española por el tema de la problemática de las pensiones dado el envejecimiento de la población actual y el incremento de la esperanza de vida. En el caso del ramo de no vida, las modalidades que mayor impacto tienen son los seguros de autos, seguido de los seguros de salud, otros seguros de no vida, siendo los seguros de decesos los que priorizan esta modalidad y, los seguros multirriesgos que presentan un valor del 11,3%, siendo en estos últimos donde se integran los seguros de hogar.

2.2.1.1. EL NEGOCIO MULTIRRIESGO

El contrato de un seguro multirriesgo se caracteriza por cubrir un conjunto de riesgos que proceden de ramos tradicionales y, a su vez, proporcionan un conjunto de garantías en una única póliza. Estos seguros destacan por la flexibilidad a la hora de la contratación ya que como principal ventaja permiten escoger las coberturas en función de las necesidades de cada tipo de asegurado. El seguro multirriesgo, a su vez, se divide en las siguientes subcategorías, siendo las más habituales:

La focalización de cobertura contra daños materiales y de responsabilidad tanto para el caso de persona física como jurídica convierte al seguro multirriesgo en un seguro versátil

para cualquier contratante integrando coberturas tanto de carácter obligatorias como opcionales.

La importancia de la falta de conocimiento del mercado donde las compañías operan implica que exista una deficiente segmentación de los asegurados dado el desconocimiento de los productos o servicios que demandan. Por esta razón, es importante conocer la evolución de las diferentes modalidades sobre todo si son productos flexibles donde se ofrecen diferentes garantías como es el caso de los seguros multirriesgos, ya que permite en cierta forma contemplar cuáles son los productos con mayor impacto y los que generan un mayor beneficio.

Además, el posicionamiento del mercado en el sector seguros dependerá no solamente del precio de la prima sino también del valor que les aporte las garantías contratadas por parte de los asegurados en función de sus necesidades reales y potenciales.

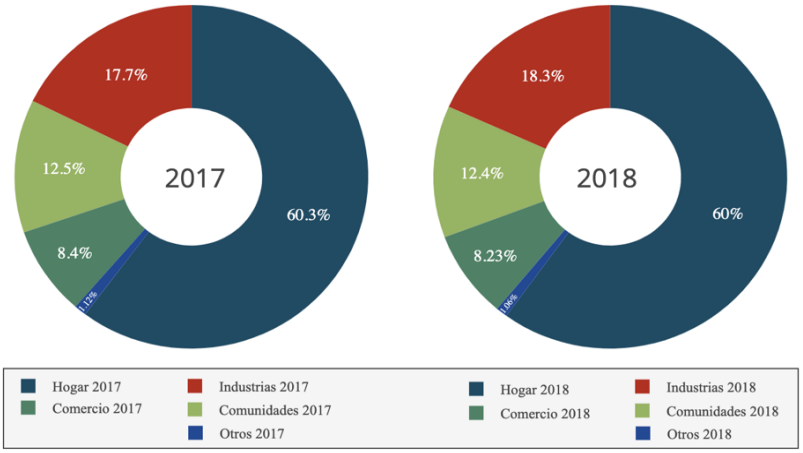


Gráfico 5. Volumen Estimado de Primas Emitidas de Seguros Multirriesgo. Elaboración Propia (ICEA, 2019)

El seguro de hogar ocupa una posición bastante ventajosa respecto al resto de subcategorías de seguros multirriesgos, obteniendo más de la mitad de la proporción en el volumen de primas emitidas. Sin embargo, la subcategoría "*Otros"* no tienen prácticamente relevancia en esta modalidad. Si se realiza la comparativa para cada uno de los periodos se observa como las puntuaciones para cada subcategoría se mantienen constantes, siendo el seguro de hogar el que lidera la primera posición en cualquiera de los casos.

Uno de los datos más relevantes a la hora de analizar cualquier sector es la tasa de crecimiento del mismo, ya que permite poder conocer la evolución experimentada de cada modalidad comprobando si se encuentra en auge o, por el contrario, en retroceso. Esto

conlleva a que se puedan adoptar medidas correctivas a tiempo y, en caso necesario, que la compañía se adapte a los cambios que el mercado experimenta, bien sea por cambios en las preferencias de los clientes como en la fijación de los precios en las tarifas.

Gráfico 6. Crecimiento Total Interanual Negocio Multirriesgo. Elaboración Propia (ICEA, 2019)

La evolución en el periodo 2018 tanto para las primas como para las pólizas es muy significativa, debido a que presenta una tendencia alcista para ambos casos, a excepción, del primer trimestre donde se aprecia una reducción en el volumen de pólizas a la mitad respecto al mismo periodo del año anterior. La consecuencia de dicho aumento se refleja en un incremento en la ganancia obtenida por la modalidad multirriesgo.

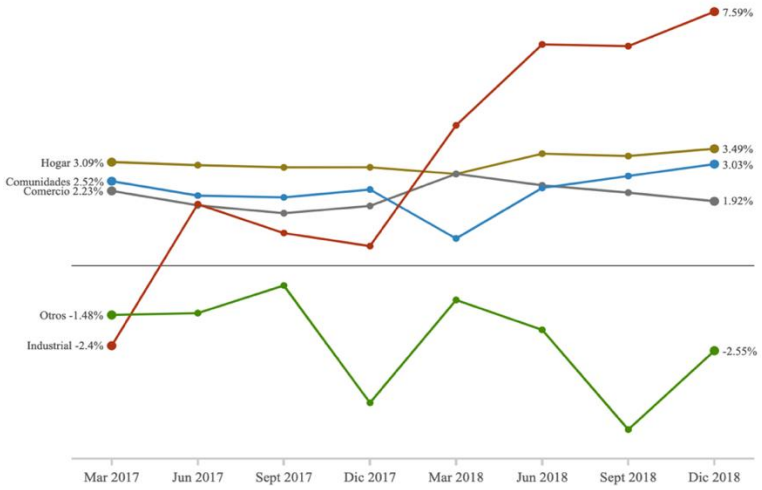


Gráfico 7. Crecimiento Interanual Primas Subcategorías Negocio Multirriesgo. Elaboración Propia (ICEA, 2019)

Al cierre del año 2018 el valor obtenido en la facturación para la modalidad de multirriesgo supuso 7.244,4 millones €, dando lugar a un incremento en la cuota anual del mercado de primas de un 4,02% y, en consecuencia, un aumento en el beneficio en el sector multirriesgo de 279,7 millones €.

Las subcategorías que mayor impacto supuso para dicho incremento fueron los seguros multirriesgos industriales con un incremento del 7,59% junto con la subcategoría de hogar que su valor ascendió a 3,49%. Dichas subcategorías supusieron 243,4 millones € del total de beneficios obtenidos por el incremento para esta modalidad de seguro.

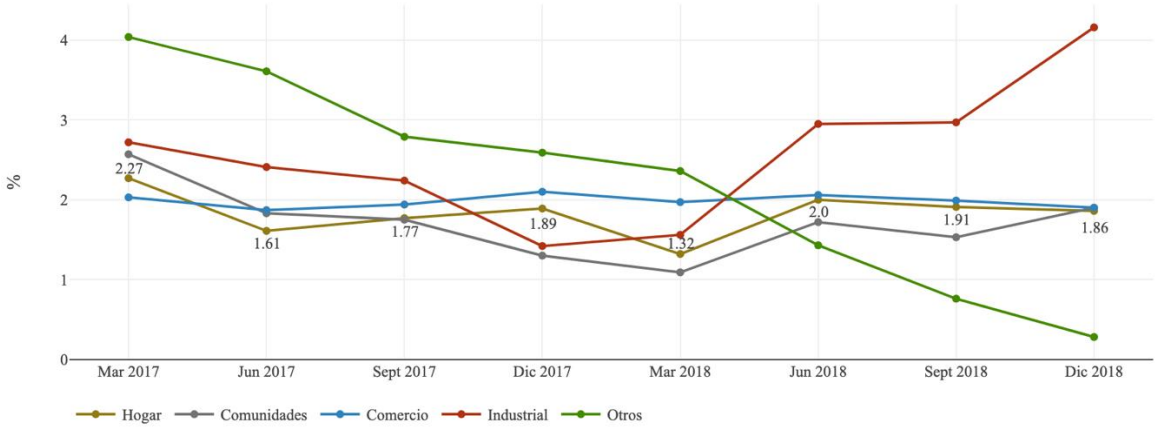


Gráfico 8. Crecimiento Interanual Pólizas Subcategorías Negocio Multirriesgo (ICEA, 2019)

Además, el crecimiento total de las pólizas en el negocio multirriesgo al cierre del 2018 supuso un incremento del 1,93%, siendo la subcategoría industrial la que mayor proporción aumentó, seguido de la modalidad de hogar que presentó un ascenso del 1,86%, igualando su valor con las modalidades de Comercio y Comunidades.

2.2.1.2. EL SEGURO MULTIRRIESGO DE HOGAR

Como se ha podido comprobar anteriormente el seguro de hogar tiene bastante relevancia en la modalidad de seguros multirriesgo. Por ello, es importante realizar una pequeña mención donde se analice este tipo de seguro de forma más detallada.

Según detalla (Wikipedia, 2019) “el seguro multirriesgo de hogar es el seguro por el que el propietario de una vivienda (o entre otras variedades, de locales comerciales) trata de cubrirse de los riesgos de que ésta sufra daños de diversa índole”.

Las coberturas que posee este tipo de seguro están asociadas principalmente a daños materiales asociadas al bien que ofrece cobertura. Sin embargo, también cubre la

responsabilidad civil que puede ocasionar perjuicio a terceras personas que se vean involucradas por el acaecimiento de un siniestro. De esta forma, algunas de las principales coberturas obligatorias asociadas a los seguros de hogar son daños por:

- Agua y electricidad
- Fenómenos atmosféricos
- Desperfectos estéticos
- Incendios
- Robo

Entre otros, siendo este último, la cobertura que se tendrá en consideración para la aplicación práctica del presente trabajo.

En la *Ilustración 3*Ilustración 3 se puede visualizar cuáles son las CCAA que en promedio más invierten en seguros de hogar anual, siendo un tipo de seguro de carácter no obligatorio.

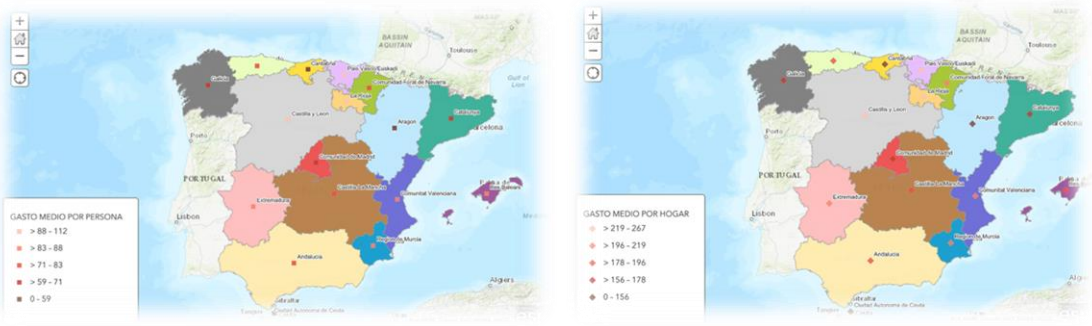


Ilustración 3. Gastos Medio en Seguros de Hogar por CCAA. Elaboración propia (INE, 2015)

A primera vista, esta comparativa permite obtener una orientación sobre la tasa de penetración que podría tener el sector asegurador en cada zona, aunque habría que tener en consideración otros factores influyentes. Se puede apreciar que el dato más relevante es el proporcionado por el País Vasco junto con Castilla-León que obtienen el rango máximo en gasto medio por viviendas aseguradas, mientras que, el puesto contrario, lo ocupa la región de Cantabria.

3. SISTEMAS DE INFORMACIÓN GEOGRÁFICA

3.1. GEORREFERENCIAR: NUEVA VISIÓN TECNÓLOGICA

Cada vez son más las compañías dedicadas al sector seguros, entre otras, que apuestan por la automatización de los procesos. El continuo dinamismo del entorno conlleva a que las compañías se tengan que adaptar para obtener una visión más completa y, a su vez, más real para la toma de decisiones que viene dada por el valor intrínseco que proporciona un profundo análisis de los datos para poder capturar los riesgos en ciertos seguros, como es el caso del seguro de hogar.

La inversión que supone utilizar técnicas que permitan un mayor conocimiento del sector como es el caso de la utilización de los sistemas de información geográfica, se suple con el beneficio que la compañía pueda obtener en un periodo a corto plazo al poder tomar decisiones con mayor certeza.

El concepto de sistemas de información geográfica, de aquí en adelante denominado con las siglas *SIG*, abarca multitud de perspectivas según diferentes autores. (Tor Bernhardsen, 2002) define este término como "un sistema vinculado por varios elementos de hardware, software y procedimientos que permiten la manipulación de datos geográficos, con el objetivo de resolver problemas complejos".

Es importante destacar que "geolocalizar" y "georreferenciar" abarcan dos términos muy semejantes, pero son dos conceptos distintos. Por ello, la *Ilustración 4* muestra a modo resumen las principales diferencias que existen entre ambos términos.

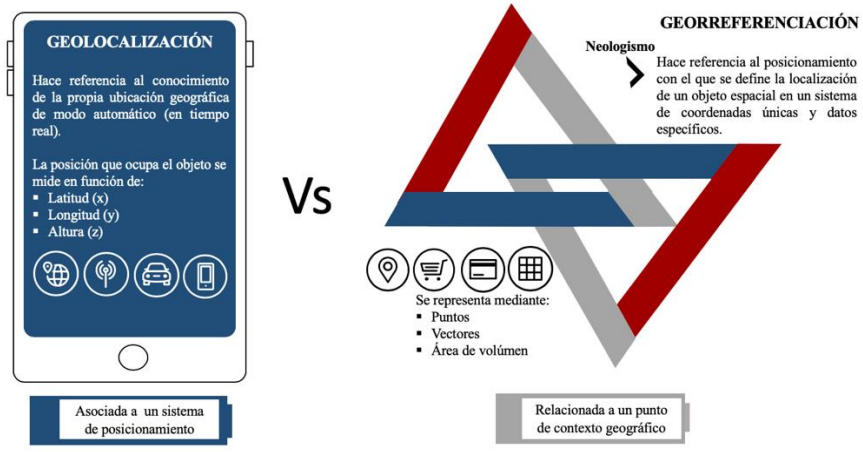


Ilustración 4. Diferenciación entre Geolocalización y Georreferenciación. Elaboración Propia

La necesidad de la utilización de los *SIG* en los seguros de hogar, entre otros, emana de la importancia de la información proporcionada por los datos que se requieren para llevar a cabo los cálculos en los modelos de tarificación de las primas. Estos sistemas permiten mayor capacidad de análisis a la hora de la modelización de la frecuencia siniestral pudiendo georreferenciar aquellas variables que presenta una mayor influencia en las zonas de la vivienda asegurada por la cobertura de robo. Por ello, aunque la aplicación de la metodología geoespacial ha sido principalmente utilizada a lo largo de los tiempos en el ámbito científico y de gestión, actualmente está ganando un gran interés por parte del sector asegurador, entre otros sectores.

Las funcionalidades de los SIG son muy variadas, siendo algunas de ellas:

- Optimización de rutas

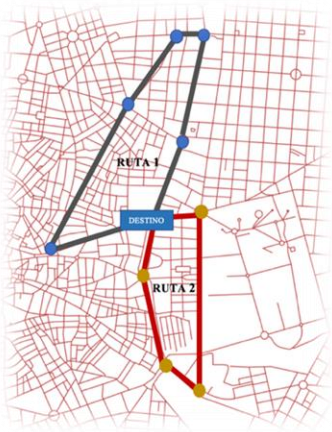
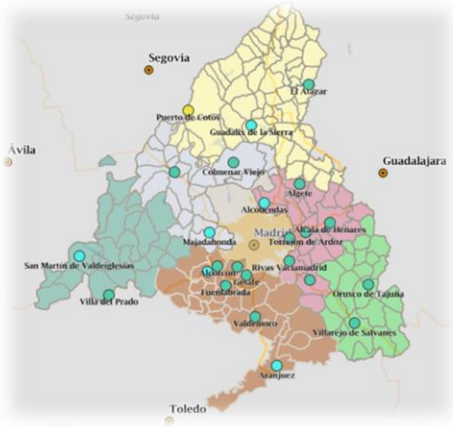


Ilustración 5. Rutas Bici Madrid. Elaboración propia

- Tendencias



El Índice reflejará el peor nivel de cualquiera de los cinco contaminantes.

CALIDAD DEL AIRE	Índice de Calidad del Aire (basado en la concentración de los contaminantes en µg/m3)				
Contaminantes considerados	**Muy bueno**	**Bueno**	**Regular**	**Malo**	**Muy malo**
Partículas menos de 2.5 µm (PM2.5)	0-10	11-20	21-25	26-50	51-800
Partículas menos de 10 µm (PM10)	0-20	21-35	36-50	51-100	101-1200
Dióxido de Nitrógeno (NO2)	0-40	41-100	101-200	201-400	401-1000
Ozono (O3)	0-80	81-120	121-180	181-240	241-600
Dióxido de Azufre (SO2)	0-100	101-200	201-350	351-500	501-1250

Ilustración 6. Calidad del Aire en Madrid. Fuente: (Comunidad de Madrid, 2019)

- Relaciones en los datos y conexiones o interrelaciones entre fenómenos

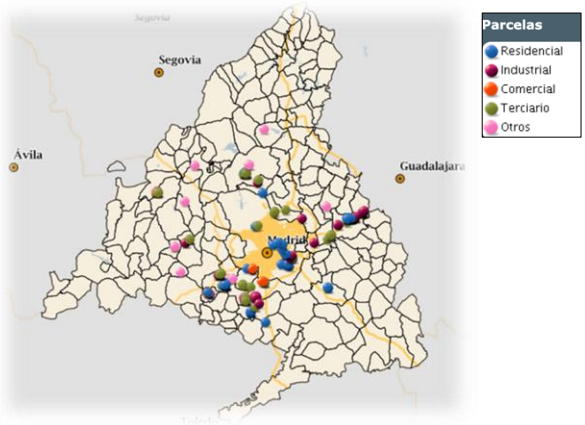


Ilustración 7. Mapa del Suelo de Madrid (Comunidad de Madrid, 2019)

- Localización.

En este caso, un ejemplo sería ¿Cómo se distribuye el turismo que se hospeda en viviendas de alquiler superiores a 750 $ mediante la plataforma *Airbnb* en los distintos barrios de Madrid? El gráfico permite observar que la zona centro de la ciudad es la preferida por los extranjeros, además, cuanto más céntrico es el barrio más renta pagan es la vivienda alquilada. En el caso de la localización, respondería a la pregunta ¿Dónde existe mayor demanda de viviendas y, en consecuencia, que el precio del alquiler/compra sea más elevado?

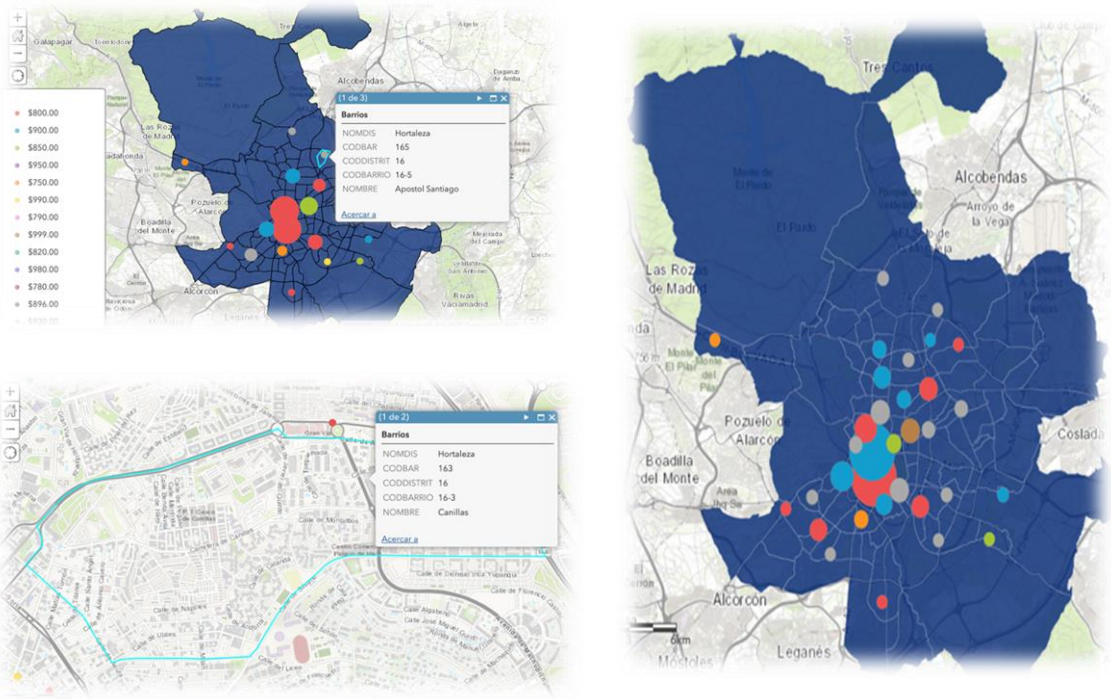


Ilustración 8. Viviendas en alquiler Airbnb con valor ≥750$. Elaboración propia

- 👁 Identificación de patrones

Aunque está fuera del contexto actuarial, un ejemplo muy curioso en el cual se aplica este tipo de técnicas es en el mundo del deporte de élite, donde las ligas profesionales de fútbol cuentan con departamentos de Data Science cuyos objetivos son analizar las habilidades de los equipos rivales, como son las estrategias que tienden a seguir según el resultado, el campo en el que juegan e incluso la hora de los partidos. Además, analizan de forma detallada los movimientos los jugadores rivales más competentes, posiciones en las que se sienten más incómodos, dónde han metido más goles, etc… mediante análisis espacial.

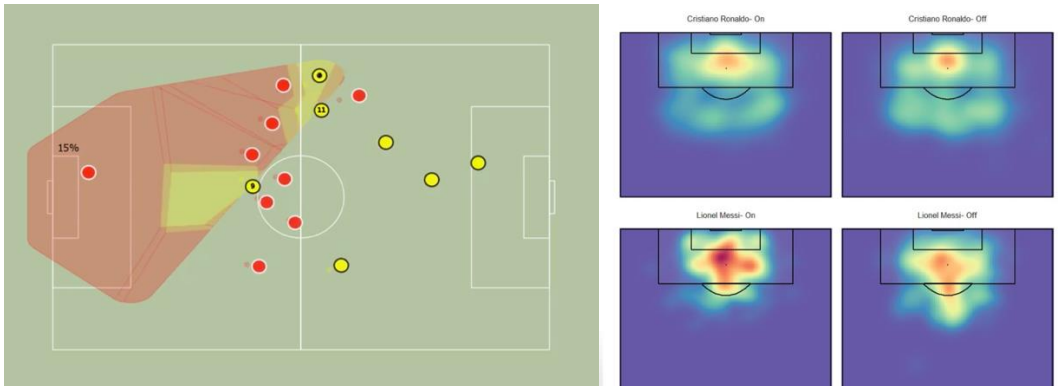


Ilustración 9. Final Argentina. HeadMap para Ronaldo y Messi. Fuente: (Ismael Gómez Schmidt, 2018)

Por un lado, la primera figura muestra la jugada de los futbolistas que contribuyeron al contraataque realizado en el partido del 9 de diciembre de 2018 en la final de Argentina entre los equipos River Plate y Boca Junior, en la cual se marcó el primer gol a favor del Boca Junior con la asistencia del jugador Nández. Por otro lado, la segunda figura muestra un mapa de calor localizado por las posiciones que ocupan los jugadores de Cristiano Ronaldo y Messi en función de su mayor o menor contribución en el partido.

- 🗺 Nuevos descubrimientos.

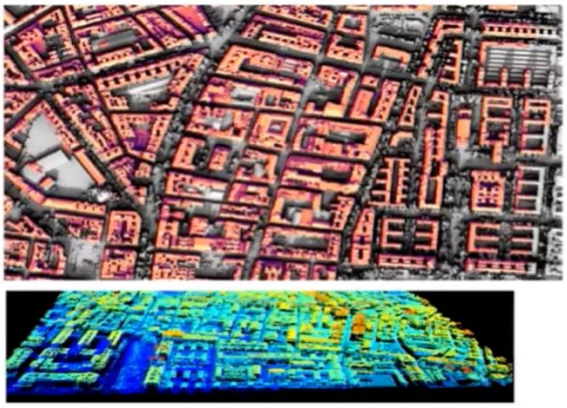


Ilustración 10. Catastro Solar en Madrid. Fuente: (Idealista. David Rey, 2019)

- Modelización.

Por otra parte, es interesante tener en cuenta las principales ventajas e inconvenientes que los *SIG* ofrecen.

Ventajas:

- ☑ Variedad de niveles o capas (SSCC, código postal, barrios, distritos, municipios, etc.)
- ☑ Almacenamiento individual de los datos y capas correspondientes, pero permite la presentación de los mismos de forma conjunta.
- ☑ Actualizaciones y posibilidad de edición de los datos.
- ☑ Uso de los datos espaciales y sus respectivas cualidades o atributos de forma paralela.
- ☑ Modelización y capacidad de análisis.
- ☑ Reducción en los costes derivado de la eficiencia de los resultados.
- ☑ Ayuda a la toma de decisiones.

Inconvenientes:

- ☒ Problemática a la hora de la conversión de los datos (de análogos a digital).
- ☒ Necesidad de factor humano especializado.
- ☒ El mantenimiento de equipos presenta un alto coste.
- ☒ Percepción de exactitud errónea.

3.2. COMPOSICIÓN DE LOS SIG

La estructura de un *SIG* está compuesta por la agrupación de un conjunto de subsistemas que interactúan entre sí. Cada elemento tiene una función esencial dentro del sistema global que forma el *SIG*, pero a su vez, todos comparten una estrecha relación ya que la inexistencia de alguno de ellos daría lugar a que los *SIG* no fueran una herramienta válida.

Los principales elementos son:

- ***Datos***. Son considerados la base principal ya que de ellos depende la información geoespacial que se requiere para poder realizar representaciones gráficas acompañadas de los atributos o cualidades en cada nivel.
- ***Software***. Es vital trabajar con una herramienta de computación que te permita realizar los análisis correspondientes con los datos de interés.

- ***Hardware***. Es necesario un conjunto de elementos físicos que lleve a cabo la implementación del software.
- ***Factor Humano***. Es el encargado de desarrollar los procedimientos.
- ***Procedimientos o métodos.*** Se requiere de metodologías consistentes que den resultados coherentes.

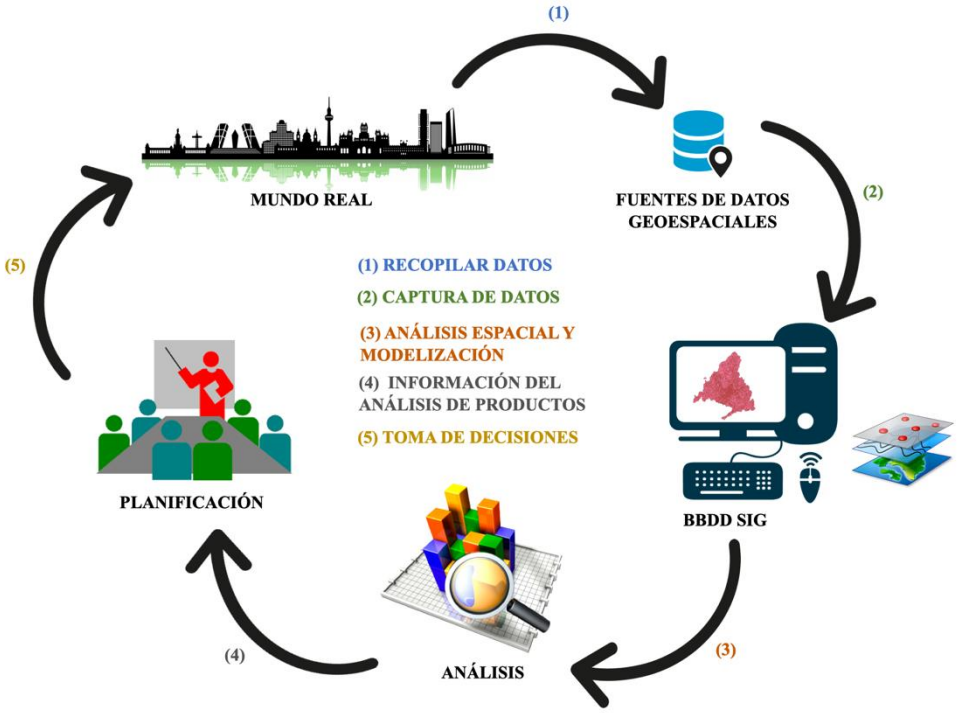


Ilustración 11. Ciclo de los SIG. Elaboración propia

3.3. ORIGEN Y EVOLUCIÓN HISTÓRICA DE LOS SIG

La introducción de los nuevos avances tecnológicos focalizados en optimizar los diferentes procesos en el entorno empresarial pone de manifiesto pensar que los sistemas de información geográfica (*SIG*) son nuevas metodologías que se están implantando a consecuencia del impacto tecnológico actual. Sin embargo, son técnicas que se han ido desarrollado desde el siglo XIX y que en la actualidad se están introduciendo en diferentes ámbitos de actuación que antaño no se tenían en consideración.

El nacimiento del análisis geoespacial se debe al geógrafo francés *Charles Picquet*, quién en el año 1832 utilizó la cartografía tradicional para representar los resultados de la contingencia de fallecimiento en los distritos de la ciudad de París a causa de la epidemia de cólera.

Ilustración 12. Cartografía (Wikipedia, 2019)

Años más tarde en 1854, el médico inglés *John Snow* realizó un trabajo similar en el que de igual forma que *Charle Picqet* representó las defunciones por cólera, pero en este caso en la ciudad de Londres. El aporte significativo que incluyó fue que además de georreferenciar las muertes producida con puntos, localizó también los pozos del agua con cruces pudiendo comprobar la zona donde más concentración de muerte existía y, en consecuencia, el foco que produjo la infección.

Ilustración 13. Cartografía (Wikipedia, 2019)

Tras estos hallazgos, el avance y la tecnología juega un papel importante para los sistemas de información geográfica, aunque a pesar del paso de los años estas técnicas no pierden su esencia basada en referenciar datos con el objetivo de solventar un problema complejo.

Los posteriores pasos relevantes que se han ido introduciendo a lo largo del tiempo hasta la actualidad en los *SIG* se detallan en la *Tabla 3* donde las etapas se desglosan mediante una escala temporal evolutiva de los diferentes periodos.

	FASE I: PERIODO DE CONCEPTUALICACION	FASE II: PERIODO DE IMPLEMENTACIÓN	FASE III: PERIODO DE MADURACIÓN	FASE IV: PERIODO DE APERTURA
PERIODOS	1975 – 1985	1985 – 1995	1995 – 1998	1998 – Actualidad
PRINCIPALES ACONTECIMIENTOS	⁂ DÉCADA DE LOS 60 **1963** – Primer **CSIG** desarrollado por Roger Tomlison (Canadá) **1964** – Aplicación **SYMAP** desarrollada por Harvard Laboratory → **Vectorial** **1969** – Aplicación **GRID** desarrollada por David Sinton en Harvard Laboratory → Ráster	⁂ DÉCADA DE LOS 80 **1985** – **GPS** (Sistema del Posicionamiento Global) **1987** – Publicación del **Periódico** Internacional de SIG **1988** – Primera **Conferencia** sobre SIG	⁂ DÉCADA MEDIADOS DE LOS 90 **1996** - Atlas digitales en línea **MultiMap** o **MapQuest** (el primer antecessor se originó en 1993 Canadá) **1997** – Aparición de **Mapserver** siendo uno de los principales servidores de catografía	⁂ SIGLO XXI **2005** – Google Maps que además de ofrecer servicios de cartografía permite desarrollar apps **2005 – Actualidad** – Formación y Desarrollo de plataformas SIG al alcance de cualquier usuario (CARTO, QGIS, Tableau, etc)
	⁂ DÉCADA DE LOS 70 **1972** – Lanzamiento del **primer satélite** de Landsat por EEUU (ERTS-1) **1973** – **MAGI** (Maryland la información Geográfica Automática) siendo uno de los primeros proyecto de SIG	⁂ DÉCADA PRINCIPIOS DE LOS 90 **1991** – **Publicación** de "SIG: Principios y Aplicaciones" por Maguire et at. 1993 – Aparación de **Xeroc PARC** como el primer servidor de mapas **1995** – Lanzamiento de **MapInfo Profesional** para Window 95		
ENFOQUE	Principalmente Cartográfico	Geo-Céntrico	Informático-Céntrico	Expansión por Tecnología

Tabla 3. Evolución de los SIG. Elaboración propia

- ***Fase I***. Está relacionada con la cartografía tradicional y, su posterior evolución para el uso de aplicaciones que se requerían almacenamiento de grandes volúmenes de datos, como catastro. Uno de los objetivos principales fue la digitalización de los mapas.
- ***Fase II***. Los SIG comienzan a utilizarse para problemas más complejos en los cuáles es necesario interrelacionar distintos niveles usando análisis espaciales y estadísticos.
- ***Fase III*.** Se orienta a toma de decisiones mediante la implementación del análisis geoespacial con herramientas cartográficas y, los resultados obtenidos por modelización.
- ***Fase IV*.** Proyección de nuevos ámbitos de actuación a consecuencia de los avances tecnológicos.

3.4. MODELIZACIÓN DE DATOS GEOESPACIALES

Como se ha comentado anteriormente, la característica fundamental de estos tipos de datos es que permiten que cualquier fenómeno u objeto pueda ser asociado a una determinada zona geográfica con una serie de atributos que los describen (datos no espaciales).

Por esta razón, aunque los términos datos e información estén estrechamente relacionados son conceptos muy distintos. Por un lado, el *dato* es simplemente el conjunto de valores que hace posible la representación del fenómeno u objeto. Sin embargo, este concepto no tiene un significado propio por sí mismo ya que requiere de una interpretación que es lo que proporciona la información para que dicho dato pueda ser utilizado con un objetivo en concreto. Por otro lado, la *información* como se ha especificado anteriormente permite al dato darle un significado propio pudiéndolo implementar de forma correcta.

Para clarificar ambos conceptos se adjunta un simple ejemplo. Los valores *40.4167* y *-3.70325* no tienen significados por sí mismo, sin embargo, si su interpretación es un código que pertenece a las coordenadas geográficas de un lugar en concreto, en este caso, a la latitud y longitud correspondientes a la ciudad de Madrid, nos proporciona una información diferente a los valores anteriores, por lo que los significados asociados a dichos valores serían: *40° 25' 0" Norte* y *3° 42' 12" Oeste*, respectivamente.

La estructura de los datos espaciales se divide en dos elementos fundamentales:

- ***Información tabulada***. Contiene la información asociada a la información geométrica.
- ***Información geométrica***. Permite la representación geoespacial mediante coordenadas, es decir, hace referencia a la posición de un fenómeno u objeto.

X	Y	0bservación	X_1
1	1	A	1
2	3	B	2
3	4	C	1
4	2	D	3

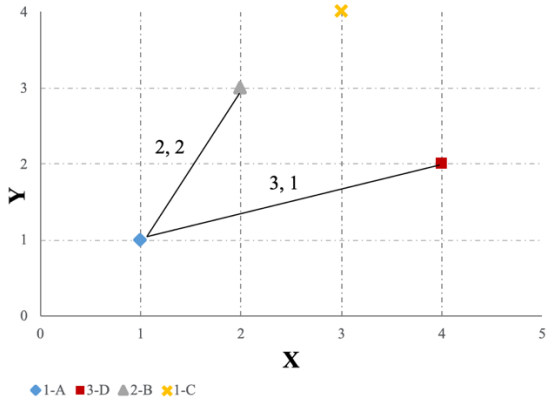


Ilustración 14. Estructura de datos espaciales. Elaboración propia

Es muy importante que a la hora de realizar la representación de cualquier fenómeno u objeto se tenga en cuenta que dependiendo del tipo de representación espacial que se aplique ésta quedará condiciona a la relación que pueda existir con otros fenómenos en el mismo espacio que se ha representado. Por esta razón, la representación de un fenómeno u objeto dependerá de la diversidad de datos geoespaciales que se le pueda aplicar.

3.4.1. INFORMACIÓN GEOESPACIAL

En el apartado anterior, se especifica que la información que se incorpora a un SIG está compuesta por dos componentes. De esta forma, la componente geométrica que permite la representación mediante coordenadas va tomar siempre un valor numérico, mientras, que en el caso de la componente tabulada podrá tomar distintos valores dependiendo de las características o atributos del fenómeno u objeto representado.

Los valores que se puede asociar a la componente tabulada son los siguientes:

- ***Valor numérico***. A su vez, se divide en las siguientes clasificaciones:

	EL VALOR NUMÉRCIO ESTABLECE:	EJEMPLO
NOMINAL	Una identificación. Es de tipo cualitativo	Número del portal de un edificio
ORDINAL	Un orden	Año de construcción de una vivienda
DE INTERVALO	Un significado por la diferencias entre valores	Renta media por hogar
DE RAZÓN	Un significado	La precipitación media de 200 mm o L/m^2 es el doble que la de 100 mm o L/m^2

- ***Valor alfanumérico*.** Hace referencia a características de los fenómenos u objetos geográficos, como por ejemplo el nombre de una ciudad.

Los cálculos numéricos que se realicen con los datos geoespaciales quedarán condicionados al tipo de variable que se utilice en su elemento tabular. De esta forma, en caso de utilizar valores nominales u ordinales, las operaciones numéricas carecerán de sentido.

3.4.2. TIPOLOGÍA DE MODELOS

Cuando se utiliza información geográfica hay que tener en cuenta que se trabaja con diferentes *dimensiones*, según (RAE, 2019) define este término como “cada una de las magnitudes que fijan la posición de un punto en un espacio”.

Cada una de las dimensiones se asocia a un tipo de formato que se utiliza para representar datos geográficos. Estos datos son implementados por dos modelos distintos: modelo vectorial y el modelo ráster.

A continuación, se detallan las dimensiones y los modelos que se utilizan en los *SIG* junto con un ejemplo ilustrativo para cada una de ellos permitiendo un mejor entendimiento de las mismas.

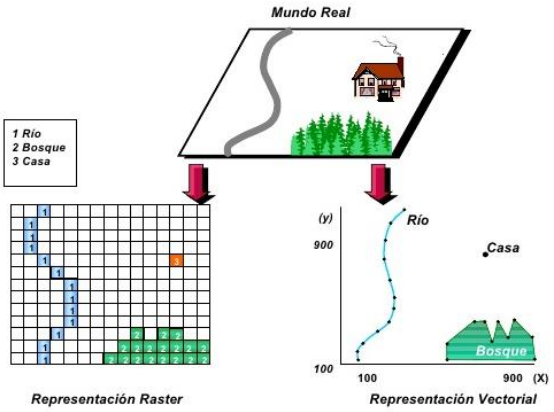


Ilustración 15. Representación Vectorial y Ráster de la Información Real

3.4.2.1. MODELO VECTORIAL

Este modelo se caracteriza por ser representado por datos con formas de puntos, líneas y polígonos.

- ◊ Puntos. Hacen referencia a los datos de cualquier fenómeno que pueda ser representado mediante unas coordenadas de longitud (X) y latitud (Y). Se caracterizan por ser adimensionales, es decir, carecen de cualidades asociadas a su localización. Ejemplo: Localización de una vivienda.

X	Y	0bservación	X_1
2	1	A	1

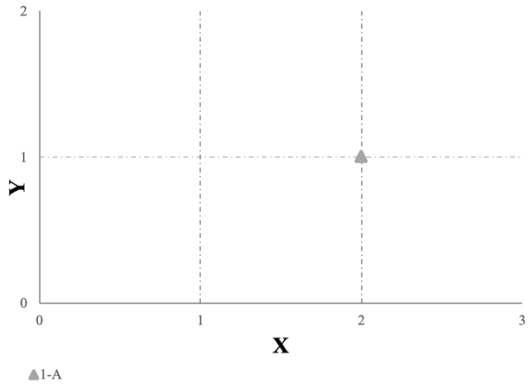


Ilustración 16. Ejemplo Puntos. Elaboración Propia

Por otro lado, se encuentran los datos que hacen referencia a cualquier objeto. Estos datos se caracterizan por ser dimensionales y, dado que solo pueden tomar valores enteros se utilizan para representar objetos cualitativos. Se clasifican en:

- – Líneas. Es la intersección entre dos puntos, mediante los cuales se puede obtener la distancia y la dirección entre los mismos. Ejemplo: Calles de una ciudad.

X_1	Y_1	X_2	Y_2	Longitud	0bservación	X_1
1	2	3	1	2,2	A	1

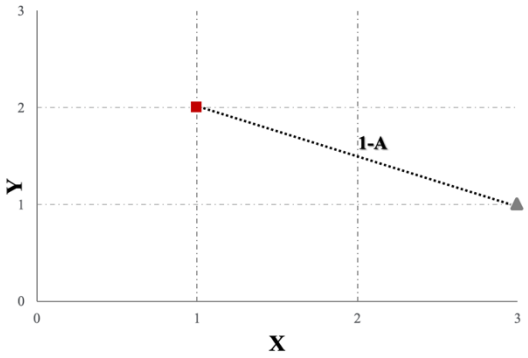


Ilustración 17. Ejemplo Líneas. Elaboración Propia

- Polígonos. Sus dimensiones son el área y el punto central (X, Y, Z). Los vértices representan sus coordenadas. Ejemplo: Barrios.

$X_{centroide}$	$Y_{centroide}$	Área	0bservación	X_1
2	1,5	2	A	1

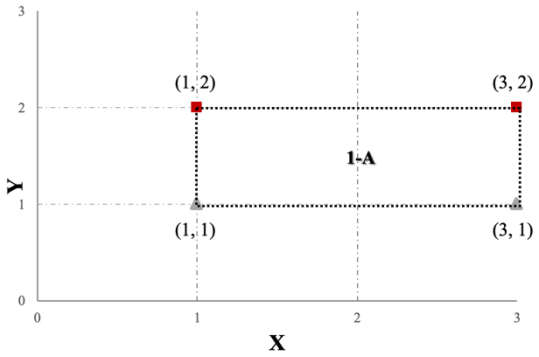


Ilustración 18. Ejemplo Polígonos. Elaboración Propia

3.4.2.2. MODELO RÁSTER

Este modelo tiene como característica fundamental que su representación se basa en el promedio de una proporción homogénea de un territorio, es decir, se codifica por celdas organizadas por filas y columnas formando una matriz. Las cuadrículas más utilizadas son:

- Polígonos de cuatro lados (rectángulos o cuadrados), siendo el caso más común.
- Polígono de tres lados (triángulos)
- Polígono de cinco lados (Hexágono)

Un punto a favor de que los elementos de las cuadrículas sean homogéneos es la posibilidad de conocer los vecinos más cercanos de cada uno, pudiendo ser medidos por la distancia euclídea o por la separación por el número de celdas.

3.4.2.3. COMPARATIVA DE MODELOS

Una vez que se ha detallado los dos tipos de modelado de datos utilizados en los SIG es interesante realizar una comparación entre ambos, ya que permite comprobar las fortalezas y debilidades que cada uno presenta con el objetivo de obtener una mejor visión de qué, cómo y por qué utilizar uno u otro.

Tabla 4. Modelización Vectorial Vs Ráster

CARACTERÍSTICAS	VECTORIAL	RÁSTER
Característica básica	Puntos, líneas y Polígonos	Trama de Pixeles
Precisión y Posición Gráfica	+	−
Cartografía Tradicional	+	−
Investigación (topología)	+	−
Capacidad de Volumen de datos	+	−
Operaciones de Cálculo	−	+
Fenómenos continuos	−	+
Propiedad de escalabilidad	+	−
Complejidad	+	−

Fuente: Elaboración Propia

3.4.2.4. CONVERSIÓN ENTRE MODELOS

Una peculiaridad que existe en utilizar datos geográficos es que es posible la conversión entre ellos. Los procesos que hacen posible este procedimiento son:

- ***Vectorización***. Siendo el caso más común, se da cuando un dato ráster es convertido en un dato vectorial. Este proceso convierte las cuadrículas en polígonos con la equivalencia correspondiente de los valores de los atributos que corresponde a una celda o pixel. La *Ilustración 19* detalla los pasos que se lleva a cabo en el proceso según (Tor Bernhardsen, 2002).

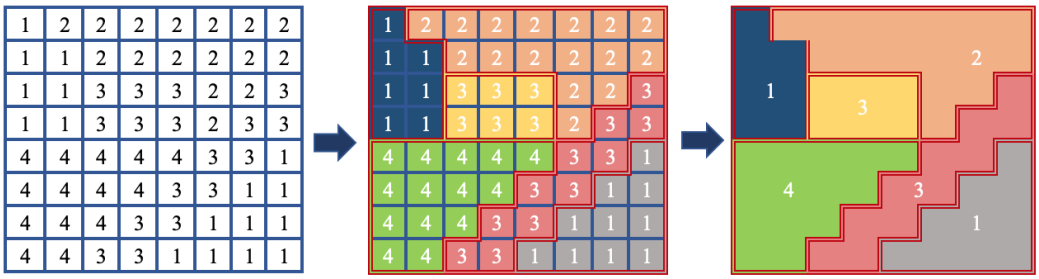


Ilustración 19. Conversión de ráster a vectorial. Elaboración propia (Tor Bernhardsen, 2002)

a. Los pixeles se asocian al valor de atributo que corresponde a un determinado polígono de forma proporcionada.
b. Se establecen los límites del polígono entre los atributos existentes.
c. El conjunto de coordenadas x e y permiten crean el polígono en función de los límites establecidos.

◊ ***Rasterización***. Es el efecto contario, es decir, cuando el dato vectorial es convertido en un dato ráster. En este caso, cada pixel es asignado a un polígono de igual valor que el atributo del polígono que corresponde. La Ilustración 15 detalla de igual forma que el procedimiento anterior los pasos que se siguen según (Tor Bernhardsen, 2002) para realizar la transformación de un tipo a otro de datos.

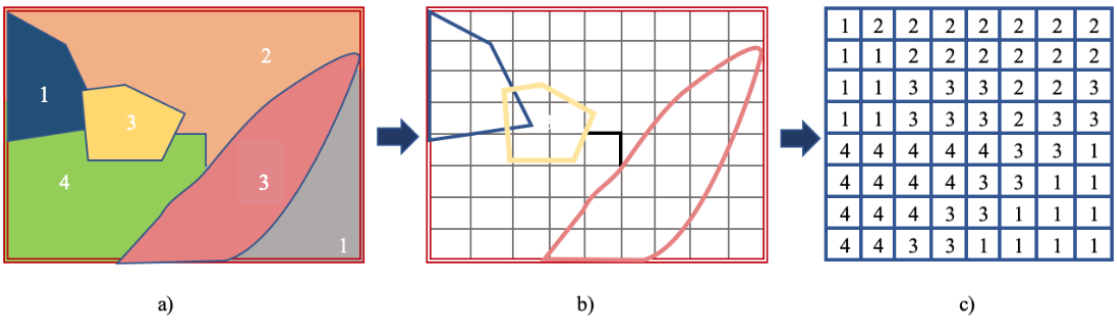


Ilustración 20. Conversión de vectorial a ráster. Elaboración Propia (Tor Bernhardsen, 2002)

a. El polígono debe de ser codificado con el valor de los atributos.
b. Se realiza cuadriculas de pixeles de igual tamaño.
c. El código de cada atributo se asocia a los pixeles que estén dentro de un mismo polígono.

Es importante remarcar que, dada la complejidad de estos procesos, la conversión de tipo de datos suele realizarse mediante un software que permita trabajar con *SIG* que ya llevan incorporado los algoritmos necesarios que permiten su implementación de forma rápida.

4. MODELOS LINEALES GENERALIZADOS (GLM)

4.1. DEFINICIÓN DE MODELO

El punto de partida del presente trabajo se basa en la modelización de la frecuencia siniestral de la cobertura de robo para los seguros de hogar para su posterior georreferenciación. Para ello, es necesario implementar un modelo que se adapte de forma adecuada a las características particulares que presentan los datos de la variable que se desee modelizar. El número de robos es una variable que presenta valores enteros no negativos, más conocidos como datos de conteo o recuento, siendo los modelos lineales generalizados (GLM) óptimos para su aplicación.

Es obvio que un modelo es una representación formal de un sistema real, con el que se pretende comprender las relaciones existentes entre las variables de interés. Pero, además, permite evaluar la posibilidad de poder controlar dichas variables y, a su vez, proporcionar un método que puede ser predictivo o explicativo en función del tipo de objetivo prefijado en el análisis.

Como menciona (Judd C. M et al., 2009) un modelo se compone fundamentalmente de dos partes:

$$\boldsymbol{DATA = MODEL + ERROR}$$

*"***DATA** representa a la respuesta o variable dependiente, que es equivalente al número de observaciones que se desea analizar descritas de una forma más compacta. **MODEL** representa la parte sistemática asociada al modelo, donde se introducen una o más variables explicativas con el objeto de dar respuesta a la variable dependiente; mientras, que la parte aleatoria está asociada al **ERROR**, que equivale a la falta de ajuste entre el modelo y la respuesta, es decir, es simplemente la cantidad por la cual el modelo no representa con exactitud los datos, siendo de gran importancia introducir este término puesto que la variabilidad en la respuesta no acaba de ser explicada*"*.

De esta forma, el proceso de análisis de datos se representa por la siguiente ecuación:

$$\boldsymbol{RESPUESTA = Parte\ Sistemática + Parte\ Aleatoria}$$

Donde: $\boldsymbol{ERROR = DATA - MODEL}$

O de forma equivalente $\boldsymbol{Parte\ Aleatoria = Respuesta - Parte\ Sistemática}$

Los Modelos Lineales Generalizados (MLG) son una extensión del modelo lineal tradicional, los cuales, permiten utilizar distribuciones no normales para realizar el modelado de cualquier variable respuesta que se ajuste a una distribución de la familia exponencial (Poisson, Binomial Negativa, Binomial, etc.), como es el caso de la frecuencia siniestral por la cobertura de robo en seguros de hogar.

4.2. ESTRUCTURA DEL MODELO

En los modelos de conteos los valores predichos (η_i) y los valores observados (μ_i) no se encuentran a la misma escala, siendo necesario un efecto multiplicativo que compense dicha desigualdad. La función de enlace utilizada es la logarítmica que permite esta transformación de forma adecuada. De esta forma, la composición general del modelado para los datos de conteos es:

Parte sistemática: $\eta_i = x_i'\beta$

Parte aleatoria: $y_i \sim Distribución\ que\ corresponda$

Función de enlace: $g(u_i) = \eta_i = x_i'\beta$

donde

η_i es el predictor lineal *(valor predicho).*

Una diferencia destacable respecto a los modelos lineales clásicos es que los supuestos canónicos de normalidad, linealidad, homocesdasticidad e independencia del modelo no son tan restrictivos, ya que no se requiere de una estructura de datos normales o la no existencia de heterocedastidad de varianzas para poder aplicarlos.

4.3. MODELOS PARA DATOS DE RECUENTO

4.3.1. MODELO DE POISSON

La distribución de Poisson se considera como el punto de partida en el modelado de datos de recuento. Una de las principales diferencias respecto a los modelos de regresión clásicos como se ha mencionado anteriormente es que la variable respuesta en el modelo de regresión de Poisson es discreta, tomando valores enteros no negativos $y = 0, 1, \ldots$ Por consiguiente, la función de probabilidad asociada al modelo es:

$$f(y) = \frac{\mu^y \cdot e^{-\mu}}{y!}, \qquad y = 0, 1, \ldots \qquad (4.\ 1)$$

donde μ es el único parámetro de la distribución, definiéndose como el promedio de ocurrencias del evento de interés por unidad de tiempo, el cual debe de ser siempre positivo (μ > 0). Al tratarse de valores discretos no negativos, los valores de *y* no tienen límite superior, en cambio, están sesgados hacia la izquierda.

Una de las características más destacadas en la distribución de Poisson es que el parámetro μ determina es su totalidad la distribución, pudiéndose demostrar de forma directa que la varianza es proporcional a la media, y ambos términos iguales al parámetro; por lo tanto, el supuesto de homocedasticidad no es adecuado para este tipo de datos (son intrínsecamente heterocedásticos), lo que define para esta distribución la propiedad de "equidispersión".

$$E(y) = Var(y) = \mu \qquad (4.\ 2)$$

Para comprender mejor el comportamiento de la distribución del modelo de Poisson para datos de recuento, se lleva a cabo una representación gráfica mostrada en la *Gráfico 9* que compara la función de masa de probabilidad para diferentes valores del parámetro μ ayudando a visualizar las características más relevantes de la distribución.

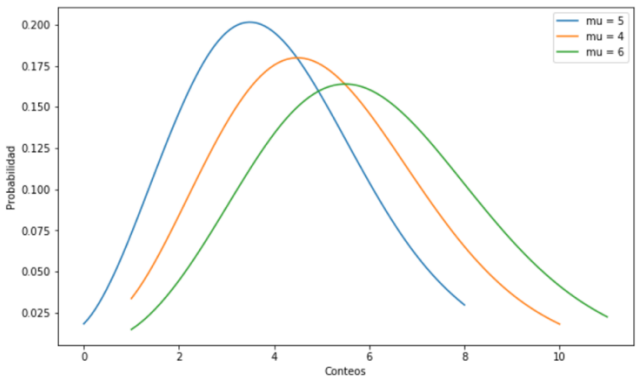


Gráfico 9. Distribución de Poisson. Elaboración Propia

Si se toma en consideración lo que se aprecia en el gráfico se puede llegar a las siguientes conclusiones:

- A medida que el valor de μ aumenta, la distribución se desplaza hacia la derecha ($E(y) = \mu$), esto es debido a que es una distribución que presenta valores enteros no negativos.
- A mayor valor de μ, mayor aproximación a una distribución gaussiana. Es el caso de la distribución de μ = 6.

- Tiene como fundamental característica que es una distribución equidispersa, es decir, $E(y) = Var(y) = \mu$. En la realidad es difícil que se de esta situación por lo que se presentan desviaciones como pueden ser:
- ☒ *Sobredispersión* (Caso muy frecuente): ***Var(y)>E(y).***
- ☒ *Infradispersión* (caso menos frecuente): ***Var(y)<E(y).***
- A medida que el el valor de μ aumenta disminuye la probabilidad en la distribución de contener recuentos observados con el valor cero.

4.3.1.1. DEFINICIÓN MATEMÁTICA Y ESTIMACIÓN DE PARÁMETROS

Si se analiza la forma matemática del modelo de regresión de Poisson su expresión adopta la siguiente forma:

$$y_i = e^{x_i'\beta} = \exp(x_i'\beta) \qquad i = 1, 2, \ldots, n \qquad (4.3)$$

donde y_i están idénticamente distribuidas como variables aleatorias Poisson con media μ_i para cada observación y $x_i'\beta$ representa el producto escalar de los regresores por el vector de los parámetros del modelo. El primer regresor que acompaña al intercepto que se incluye en el modelo tiene el valor 1.

Este tipo de modelo deriva a partir de la denominada ***función enlace*** de los Modelos Lineales generalizados, que tiene como requisito fundamental que la distribución de la variable dependiente pertenezca a la familia exponencial como se ha mencionado anteriormente, parametrizando la relación entre la media de la variable respuesta (μ_i) y el predictor lineal (x_i), adoptando la siguiente expresión:

$$\log(\mu_i) = \eta_i = x_i'\beta = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \qquad i = 1, 2, \ldots, n \qquad (4.4)$$

donde

Parte sistemática: $\eta_i = x_i'\beta$

Parte aleatoria: $y_i \sim Poisson(\mu_i)$

Función de enlace estándar: $g(\mu_i) = \eta_i = \log(\mu_i)$

Por tanto, $\mu_i = \exp(\eta_i) = \exp(x_i'\beta)$.

El uso de la función exponencial asegura que el lado derecho en la ecuación anterior sea siempre un valor positivo, como es el valor esperado de la variable de recuento y_i en el lado izquierdo de la ecuación.

El método más utilizado para la estimación de los parámetros en el modelo de regresión de Poisson es la estimación puntual por máxima verosimilitud (*ML=maximum likelihood*). Este método, tiene como propiedad maximizar el valor de la probabilidad de los datos observados debido a que se selecciona el valor del parámetro como estimador.

La función que se utiliza para el modelo de regresión de Poisson es equivalente a la maximización logarítmica (*loglikelihood*). La *función log-verosimilitud* (*MLE*) para este tipo de modelos toma la siguiente forma (Cameron, A. C and Trivedy, K.P, 1998):

$$\log L(\beta; Y, X) = \log \prod_{i=1}^{n} f(y_i/x_i\,;\beta) = \sum_{i=1}^{n} \log[f(y_i/x_i;\beta)] =$$

$$\sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_i'\beta - \log(y_i!) \qquad (4.\ 5)$$

El valor de maximización para β, denotado como $\hat{\beta}$, es la solución a *k* ecuaciones no lineales correspondientes a la condición de primer orden para la función log-verosimilitud (*MLE*) igualada a cero, de tal forma que:

$$\sum_{i=1}^{n}[y_i - \exp(x_i'\beta)]x_i = 0 \qquad (4.\ 6)$$

Si x_i incluye un término constante, las condiciones de primer orden implican que los $\sum_{i=1}^{n} \hat{u}_i = 0$, donde $\hat{u}_i$ es un residuo definido como $\hat{u}_i = y_i - \hat{E}(y_i/x_i) = y_i - \exp(x_i'\hat{\beta})$.

Los estimadores por máximo verosimilitud en el modelo de Poisson produce estimaciones únicas en los parámetros debido a que la función log-verosimilitud es cóncava. Aunque no se entrará en detalle, es conveniente mencionar que las ecuaciones de verosimilitud no son lineales en los parámetros por lo que requieren métodos numéricos de re-ajuste iterativos como el de *Newton-Raphson* para resolverlos.

Por un lado, la teoría estándar de máxima verosimilitud de modelos correctamente especificados el estimador $\hat{\beta}_j$ es consistente para β y asintóticamente normal para la matriz de covarianzas, donde;

$$Var_{ML}[\hat{\beta}_j] = (\sum_{i=1}^{n} \mu_i x_i x_i')^{-1} \qquad (4.\ 7)$$

Por otro lado, cuando los modelos no están correctamente especificados se pueden utilizar diferentes métodos alternativos para su modelización:

i. White (1982) considera una situación en la que el modelo no está correctamente especificado, así no existe ningún β tal que:

$$\sum_{i=1}^{n} \log[f(y_i/x_i;\beta)] = \sum_{i=1}^{n} \log[f_0(y_i/x_i)] \qquad (4.\ 8)$$

donde;

f corresponde a la función de densidad del modelo especificado.

$f_0(y_i/x_i)$ corresponde a la función de densidad verdadera del modelo.

Propone otro método alternativo de estimación denominado cuasi máxima verosimilitud (QML = pseudo maximum likelihood estimation). La función que se utiliza para la estimación de los parámetros es la maximización logarítmica de verosimilitud mal especificada (Wilkelman, R, 2008)

i. Se puede utilizar diferentes tipos de modelizaciones como es el caso del modelo de regresión binomial negativo, entre otros, siendo éste junto con el modelo de regresión de Poisson, uno de los más utilizado para datos de recuento.

4.3.2. MODELO BINOMIAL NEGATIVO

La definición más común de la distribución binomial negativa es el número de fracasos antes de obtener el primer éxito en diferentes pruebas realizadas. La variable dependiente del modelo es también considerada un conteo como en el caso del modelo de Poisson, por lo que toma valores enteros no negativos $y_i = 0, 1, \ldots$

El modelo presenta la siguiente función de probabilidad:

$$\mathrm{P}(Y = y) = \binom{y + r - 1}{y} p^r (1-p)^y \quad y = 0, 1, \ldots; \text{ y } 0 \leq p \leq 1 \qquad (4.\ 9)$$

donde *y* es el número de fracasos antes del r-ésimo éxito en una prueba o experimento, *p* la probabilidad de que ocurra un suceso y $q = 1-p$ la probabilidad de fracaso.

$$\mathrm{E(y)} = \mathrm{r}\frac{(1-\mathrm{p})}{\mathrm{p}} \qquad \mathrm{Var(y)} = \mathrm{r}\frac{(1-\mathrm{p})}{\mathrm{p}^2} \qquad (4.\ 10)$$

La peculiaridad que presenta la distribución binomial negativa es que cuando se da el caso de $r = 1$ es equivalente a una distribución geométrica de parámetro *p*, definida como el número de fracasos antes del primer éxito. La expresión que adopta es la siguiente:

$$\mathrm{P(Y = y)} = \binom{\mathrm{y} + 1 - 1}{\mathrm{y}} \mathrm{p}^1 (1-\mathrm{p})^{\mathrm{y}} = \mathrm{p}(1-\mathrm{p})^{\mathrm{y}} \qquad (4.\ 11)$$

donde;

$$E(y) = \frac{(1-p)}{p}$$

$$Var(y) = \frac{(1-p)}{p^2}.$$

La mejor forma de comprender el comportamiento de esta distribución es representándola gráficamente. Por ello, la *Gráfico 10* función masa de probabilidad de los parámetros de la distribución asociados a distintos valores. En cada caso se mantiene fijo uno de los parámetros, mientras que el otro varía.

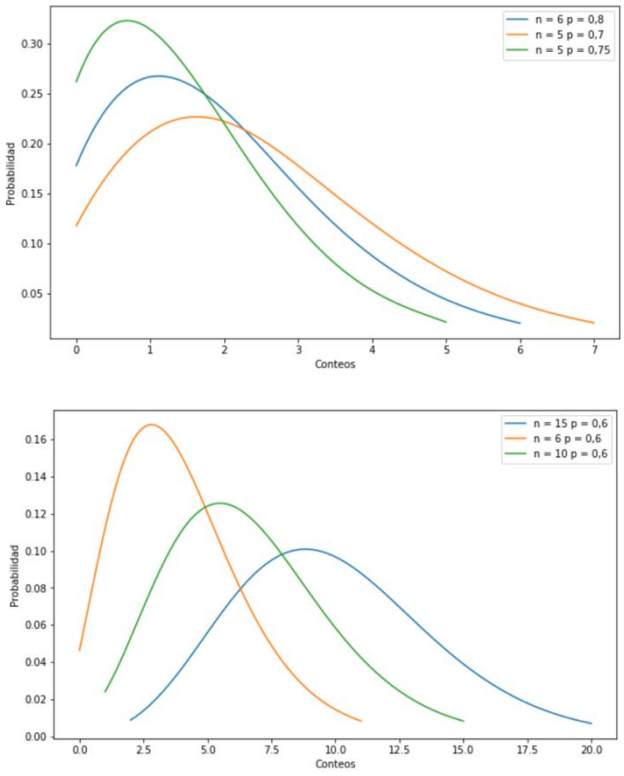


Gráfico 10. Distribución Binomial Negativa. Elaboración Propia

Otra característica que aporta la distribución binomial negativa que a comparación de la distribución de Poisson en caso de existencia de *sobredispersión*, este modelo aporta resultados más favorables ya que el parámetro *r* de la distribución tiende a controlarla. En el caso $r = 0$ distribución se reduce a una Poisson con $Var(y) = \mu + 0 \cdot \mu = \mu$.

4.3.2.1. DEFINICIÓN MATEMÁTICA Y ESTIMACIÓN DE LOS PARÁMETROS

Si se analiza la forma matemática más común del modelo de regresión binomial negativo es una mixtura entre una distribución Poisson y una distribución gamma. Esto es

consecuencia de que como anteriormente se ha mencionado, en el modelo de regresión de Poisson el parámetro ($\mu_i = \lambda$) se considera constante para cada nivel de las covariables, es decir, para individuos con las mismas características que determinan las variables (*xi*) del modelo la media es la misma, siendo difícil que en la práctica suceda, lo que da lugar a problemas de heterogeneidad (sobredispersión). Por este motivo, el modelo de regresión binomial negativo es considerado como un modelo compuesto ((Wilkelman, R, 2008), entre otros) debido a que la variable respuesta *(y)* se distribuye por una distribución de Poisson y la media μ es considerada como una variable aleatoria que se distribuye como una distribución gamma (Hilbe, J.M, 2011).

El modelo binomial negativo presenta diferentes parametrizaciones, siendo la parametrización binomial negativa de tipo II (comúnmente más conocida como *NegBin II*) la que se tendrá en consideración en el presente trabajo, ya que es considerado el caso más extendido.

El modelo de regresión binomial negativo puede ser expresado:

$$f(y_i|\mu_i,\alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \qquad y_i = 0,1,\ldots \quad (4.\ 12)$$

$$\alpha \geq 0$$

donde α e y_i son los parámetros de la distribución y $\Gamma(\cdot)$ es la función gamma. El parámetro α determina el grado de dispersión en las predicciones.

$$\mathrm{E}(\mathrm{y_i}) = \mu_i \qquad \mathrm{Var}(\mathrm{y_i}) = \mu_i + \alpha\mu_i^2 \quad (4.\ 13)$$

donde;

- ***Parte sistemática:*** $\eta_i = x_i'\beta$
- ***Parte aleatoria:*** $y_i \sim BN(\alpha, \mu_i)$
- ***Función de enlace estándar:*** $g(\mu_i) = \eta_i = \log(\mu_i)$

Por tanto, $\mu_i = \exp(\eta_i) = \exp(x_i'\beta)$.

El método utilizado para la estimación de los parámetros es el mismo que para el caso del modelo de Poisson, siendo este la estimación puntual por máxima verosimilitud (ML = maximum likelihood). Para este tipo de modelo la *función log-verosimilitud* (MLE = log-likelihood). (Cameron, A. C and Trivedy, K.P, 1998)

$$\log L(\beta,\alpha) = \sum_{i=1}^{n}\left\{\begin{array}{c}\left(\sum_{j=0}^{y_i-1}\log(j+\alpha^{-1})\right) - \log y_i! - (y_i+\alpha^{-1})\log\left(1+\alpha\exp(x_i'\beta)\right) \\ +y_i\log\alpha + y_i x_i'\beta\end{array}\right\} \quad (4.\ 14)$$

A través del método de máxima verosimilitud en el caso del modelo NegBin II, se obtienen estimadores consistentes tanto para el parámetro del modelo β como para el parámetro de dispersión α.

La condición de primer orden para la función log-verosimilitud (MLE) igualada a cero se corresponde con el valor de maximización para β, denotado como $\hat{\beta}$; y el valor de maximización para α, denotado como $\hat{\alpha}$, dando lugar a que:

$$\sum_{i=1}^{n}\left(x_i\frac{y_i-\mu_i}{1+\alpha\mu_i}\right)=0 \qquad (4.\ 15)$$

$$\sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}\left(\log(1+\alpha\mu_i)-\sum_{j=0}^{y_i-1}\frac{1}{(j+\alpha^{-1})}\right)+\frac{y_i-\mu_i}{\alpha(1+\alpha\mu_i)}\right\}=0 \qquad (4.\ 16)$$

Los métodos numéricos de *Newton-Raphson* también forman parte de la construcción de los estimadores por el método de máxima verosimilud, como en el caso de Poisson.

Cuando el modelo está especificado correctamente, existe consistencia en los estimadores de $\hat{\beta}$ para β y de $\hat{\alpha}$ para α y, por tanto, considerándose asintóticamente normales las matrices de covarianzas. Siendo,

$$Var_{ML}[\hat{\beta}_{NBII}]=\left(\sum_{i=1}^{n}\frac{\mu_i}{1+\alpha\mu_i}x_i x_i'\right)^{-1} \qquad (4.\ 17)$$

$$Var_{ML}[\hat{\alpha}_{NBII}]=\left(\sum_{i=1}^{n}\frac{1}{\alpha^4}\left(\log(1+\alpha\mu_i)-\sum_{j=0}^{y_i-1}\frac{1}{(j+\alpha^{-1})}\right)^2+\frac{\mu_i}{\alpha^2(1+\alpha\mu_i)}\right)^{-1} \qquad (4.\ 18)$$

y, $Cov_{ML}[\hat{\beta}_{NBII},\hat{\alpha}_{NBII}]=0$.

4.3.3. PRINCIPALES MEDIDAS DE BONDAD DE AJUSTE GLM

Este apartado pretende mencionar las principales medidas utilizadas para comprobar si el modelo de conteo se encuentra correctamente especificado. Puesto que implementar un modelo que permita explicar la frecuencia de robo para los seguros de hogar es una de las fases del estudio, pero no es el objetivo principal, siendo éste, medir el potencial geográfico en las zonas de interés, este apartado se detallará solamente a modo resumen.

Tabla 5. Medidas de Bondad de Ajuste GLM

MEDIDA	EXPRESIONES GENERALIZADAS
DEVIANCE	$D(y_i, \hat{\mu}_i) = -2\{\mathcal{L}(y_i) - \mathcal{L}(\hat{\mu}_i)\}$ (4. 19) donde; ▪ $\mathcal{L}(y_i)$ equivale a la función log-verosimilitud del modelo completo (modelo de *k* parámetros, uno por observación). ▪ $\mathcal{L}(\hat{\mu}_i)$ equivale a la función log-verosimilitud del modelo actual.
COEFIENTE DE DETERMINACIÓN BASADO EN LA DEVINACE	$R^2_{Deviance} = 1 - \left(\frac{\text{Deviance Residual}}{\text{Deviance Nula}}\right)$ (4.20) $0 < R^2_{DEVIANCE} < 1$ ▪ La deviance nula hace referencia a la desviación del modelo que no depende de ninguna variable explicativa introducida en el mismo, es decir, solo tiene en cuenta el término constante. ▪ La deviance residual indica la desviación que depende de las diferentes variables explicativas que han sido introducidas en el modelo.

Además, de estas métricas existen otras medidas de bondad de ajuste que permiten realizar comparación entre modelos siendo el criterio de Información Akaike (AIC) y el criterio de Información Bayesiano (BIC), los más utilizados en la práctica. Siendo:

$$BIC = k * \ln(n) - 2 * \ln(L) \quad (4.21)$$

$$AIC = 2k - 2 * \ln(L) \quad (4.22)$$

Siendo k el número de parámetros y $\ln(L)$ la función log-verosimilitud del modelo utilizado.

5. ANÁLISIS DE RESULTADOS

5.1. SELECCIÓN Y MINERÍA DE DATOS

La fase de recolecta de los datos ha sido sin duda una parte bastante compleja, debido a la falta de información pública relacionada con el nivel de distritos, en concreto, para la comunidad de Madrid. Tras una larga investigación en multitud de fuentes, se ha conseguido un histórico mensual de 4 años perteneciente al periodo 2015-2018, ambos inclusive.

Las variables que son parte del estudio seleccionadas finalmente con una frecuencia mensual en el presente trabajo se adjuntan en el

Anexo B**.**

Dado la dificultad de encontrar datos mensuales, en algunos de los casos, se han tenido que establecer hipótesis que permitan convertir variables con un periodo trimestral en variables con un periodo mensual. Para mayor detalle, se especifica los cálculos realizados en cada una de las hipótesis el año 2018, ya que en todos los casos se ha llevado a cabo el mismo proceso.

HIPÓTESIS 1. Densidad de Población

Esta variable presenta los habitantes por hectáreas de cada uno de los distritos anualmente. Para realizar el cambio se ha tenido en cuenta que en todos los meses del mismo año la superficie por hectárea se mantiene constante, y se ha recalculado la variable teniendo en cuenta el total de la población mensual de cada distrito. Para ello, se ha utilizado la siguiente fórmula:

$$\text{Densidad de Población} = \frac{\text{Población Total}}{\text{Superficie Ha}} \qquad (5.\ 1)$$

Tabla 6. Cálculo Densidad de Población Mensual por Distritos

DISTRITOS	AÑO	MES	Superficie_Ha	Población	Densidad_Ha_Hab
Centro	2018	Diciembre	522,8246211	132.178	253

Fuente: Elaboración Propia

HIPÓTESIS 2. Número de Robos en Viviendas

Los datos que se adjuntan en la *Tabla 7* no presentan de forma específica cuántos de los robos relacionados con el patrimonio pertenecen a la categoría de hogar.

Tabla 7. Criminalidad por Distritos

DISTRITOS	RELACIONADAS CON LAS PERSONAS	RELACIONADAS CON EL PATRIMONIO	POR TENENCIA DE ARMAS	POR TENENCIA DE DROGAS	POR CONSUMO DE DROGAS
CENTRO	26	62	219	193	80
ARGANZUELA	8	8	1	2	0
RETIRO	3	5	2	5	0
SALAMANCA	6	51	3	33	4
CHAMARTÍN	8	19	2	30	6
TETUÁN	24	23	1	16	4
CHAMBERÍ	10	9	0	23	7
FUENCARRAL - EL PARDO	5	2	0	43	3
MONCLOA - ARAVACA	14	21	0	9	2
LATINA	11	10	4	5	2
CARABANCHEL	15	11	2	5	0
USERA	5	8	6	11	3
PUENTE DE VALLECAS	14	24	10	4	5
MORATALAZ	3	1	0	14	1
CIUDAD LINEAL	13	9	0	8	0
HORTALEZA	11	10	3	20	5
VILLAVERDE	17	34	2	23	1
VILLA DE VALLECAS	5	9	1	12	17
VICÁLVARO	1	0	2	3	0
SAN BLAS - CANILLEJAS	13	17	1	31	1
BARAJAS	3	13	0	1	5
SIN DISTRITO ASIGNADO	275	327	16	46	63
TOTAL	**490**	**673**	**275**	**537**	**209**

Fuente: (Ayuntamiento de Madrid, 2019)

Por ello, se ha escogido una segunda base de datos que representa el total poblacional, es decir, el municipio de Madrid con todos sus distritos especificando cuál es la proporción que corresponde a robos en viviendas.

Tabla 8. Criminalidad de Madrid por Robos en Viviendas

MUNICIPIO MADRID	T1	T2	T3	T4	T1 (No Acum.)	T2 (No Acum.)	T3 (No Acum.)	T4 (No Acum.)
Robos con Fuerza en Domicilios, Establecimientos y Otras Instalaciones	2.262	4.374	6.754	5.462	1.410	1.259	1.595	1.198
Robos con Fuerza en Domicilios	1.410	2.669	4.264	5.462	1.410	1.259	1.595	1.198
Ponderación Mensual	**470**	**889**	**1.421**	**1.820**	**470**	**419**	**531**	**399**

Fuente: (Ministerio del Interior, 2019)

Con ello, lo que se ha tenido en cuenta es que en cada uno de los meses que pertenece a un mismo trimestre la frecuencia siniestral de robos se mantiene constante. De esta forma,

se puede calcular la proporción para cada uno de los distritos mediante la siguiente fórmula:

$$\text{Proporción} = \frac{\text{Robos Patrimonio}}{\text{Robos Total}} \qquad (5.\ 2)$$

Los resultados obtenidos son:

Tabla 9. Resultado Proporción/Distritos

DISTRITOS	RELACIONADAS CON EL PATRIMONIO	Proporción
CENTRO	62	0,092124814
ARGANZUELA	8	0,011887073
RETIRO	5	0,007429421
SALAMANCA	51	0,075780089
CHAMARTÍN	19	0,028231798
TETUÁN	23	0,034175334
CHAMBERÍ	9	0,013372957
FUENCARRAL - EL PARDO	2	0,002971768
MONCLOA - ARAVACA	21	0,031203566
LATINA	10	0,014858841
CARABANCHEL	11	0,016344725
USERA	8	0,011887073
PUENTE DE VALLECAS	24	0,035661218
MORATALAZ	1	0,001485884
CIUDAD LINEAL	9	0,013372957
HORTALEZA	10	0,014858841
VILLAVERDE	34	0,050520059
VILLA DE VALLECAS	9	0,013372957
VICÁLVARO	0	0
SAN BLAS - CANILLEJAS	17	0,02526003
BARAJAS	13	0,019316493
SIN DISTRITO ASIGNADO	327	0,485884101
TOTAL	**673**	**1**

Fuente: Elaboración Propia

Por último, una vez que se obtiene la proporción por distritos se extrapola a la población total, donde el resultado final se obtiene mediante el siguiente producto.

$$\text{Nº Robos en Viviendas} = \text{Proporción} * \text{Robos Madrid} \qquad (5.\ 3)$$

Tabla 10. Extrapolación Frecuencia de Robos en Viviendas por Distritos

DISTRITOS	AÑO	MES	Robos_General	Proporción	Robos_Madrid_Domicilios	Cobertura_robos_Domicilios
Centro	2018	Diciembre	62	0,092124814	399	36

Fuente: Elaboración Propia

HIPÓTESIS 3. Renta Media Alquiler (€/m^2)

El aumento o disminución en el precio del alquiler de la vivienda depende del índice de precios de consumo (IPC). La evolución de este índice permite ser medido de forma

mensual. Por ello, la renta media de alquiler por distrito se mantiene constante para los meses que corresponda a cada trimestre, pero se pondera con el IPC perteneciente la comunidad de Madrid con el mes que le corresponda.

$$\text{Renta Media Alquiler} = Renta\ Alquiler * \left(1 + \left(\frac{IPC\ Mensual}{100}\right)\right) \qquad (5.\ 4)$$

Tabla 11. Cálculo Renta Media Mensual por Distritos

DISTRITOS	AÑO	MES	Renta_Media_ Alquiler	IPC_Mensual_ Madrid	Renta_Media_ Alquiler
Centro	2018	Diciembre	19,3	-0,30	19,251

Fuente: Elaboración Propia

HIPÓTESIS 4. Precio Vivienda de 2ª Mano (€/m^2)

El incremento o decremento en el precio de compraventa de la vivienda de segunda mano depende del índice de precios de la vivienda (IPV), en este caso, perteneciente a la Comunidad de Madrid. La evolución de este índice se computa en periodos trimestrales, de igual forma, que la variable de interés. Por ello, en este caso, la variable precio de compraventa se mantendrá constante para los meses correspondientes dentro de un mismo trimestre, ya que el aplicarle la ponderación con el IPV la única variación sería el aumento o disminución del índice puesto que los precios de cada mes dentro de un mismo trimestre serían de igual valor.

HIPÓTESIS 5. Variables Estacionales: 1, 2, 3

La frecuencia siniestral de los robos en las viviendas está muy relacionada con la fecha en la que se producen los siniestros. Según una noticia del (Períodico ABC, 2016) en la mayoría de los casos los ladrones prefieren periodos en los que los inquilinos se encuentra hospedados fuera de casa, principalmente en invierno, donde la luz del sol se apaga a media tarde. Por esta razón, los meses más propensos a sufrir un robo en el hogar se da en marzo o abril (dependiendo de cada año), época de semana santa; agosto, vacaciones de verano y, diciembre, época navideña.

Dado que la fecha es una variable relevante, se lleva a cabo la realización de tres variables ficticias donde se codifica con 1 si la fecha es alguna de los meses propensos a robos y 0 en caso contrario.

- Estacional 1

$$\begin{cases} 0 & \text{No Semana Santa} \\ 1 & \text{Semana Santa} \end{cases}$$

- Estacional 2

$$\begin{cases} 0 & \text{No Navidad} \\ 1 & \text{Navidad} \end{cases}$$

- Estacional 3

$$\begin{cases} 0 & \text{No Vacaciones} \\ 1 & \text{Vacaciones} \end{cases}$$

Por último, se ha llevado a cabo la creación de una nueva variable que contabiliza el *número de días festivos* en calendario laboral.

HIPÓTESIS 6. Tasa de Robos en Viviendas

Esta variable es de vital importancia debido a que se va a utilizar como variable dependiente en la modelización realizada. Al no disponer de una base de datos real cedida por una compañía, la variable de la tasa o probabilidad de robos en viviendas se debe de aproximar con los datos de los que se dispone, ya que la tarifa de un seguro de hogar se calcula combinando dos modelos diferentes: uno para la cuantía y otro para la frecuencia siniestral, siendo este último el que se realizará utilizando únicamente variables exógenas al riesgo.

Normalmente, para realizar el cálculo de la frecuencia siniestral se utiliza el número de robos relativizado por la exposición al riesgo, dando lugar a una tasa. En este caso, se procede a calcular la tasa que será escalada para poder obtener valores enteros no negativos y poder ser modelizada mediante el modelo que corresponda.

Para su cálculo, se toma en consideración el valor del tamaño medio del hogar perteneciente a cada uno de los años analizados de la Encuesta continua de Hogares publicada por el (INE, 2019). Los cálculos que se utilizan son los siguientes:

$$\text{Nº Viviendas/Distritos} = \frac{\text{Población Total}}{\text{Tamaño Medio Hogar}} \qquad (5.\ 5)$$

$$\text{Tasa o Probabilidad de Robo} = \frac{\text{Nº Robos}}{\text{Nº Viviendas}} \qquad (5.\ 6)$$

Donde el valor promedio del tamaño medio del hogar es 2,5 personas/vivienda.

5.2. ANÁLISIS EXPLORATORIO GEORREFERENCIADO

En este apartado se incluye un análisis preliminar de las variables que se consideran más interesantes para el estudio. Normalmente, el análisis descriptivo permite comprobar a priori el comportamiento que sigue cada una de las variables con el objetivo de visualizar la existencia de posibles valores atípicos, además de posibles patrones que puedan influir posteriormente en la modelización de los datos.

En primer lugar, se adjunta para el caso de las variables numéricas un resumen de los valores que corresponden a medidas de tendencia central (media y cuartiles) y dispersión (desviación típica y rango (Max - Min)).

Tabla 12. Descriptivo Variables Cuantitativas

	Periodo	Dias_Festivos_Laborales	Pob_Extranjera_Total	Superficie_Ha_Mes	Densidad_Poblacion	Paro
count	1008.000000	1008.000000	1008.000000	1008.000000	1008.000000	1008.000000
mean	2016.500000	1.083333	19251.315476	2878.147447	140.274682	9204.121032
std	1.118589	0.975876	10029.517823	4891.333845	94.714445	4709.088400
min	2015.000000	0.000000	3891.000000	467.918499	9.835902	1741.000000
25%	2015.750000	0.000000	11618.500000	610.319194	65.117602	5455.750000
50%	2016.500000	1.000000	16382.500000	1404.835796	155.145661	8529.500000
75%	2017.250000	2.000000	27153.000000	2741.977134	218.086951	11385.250000
max	2018.000000	3.000000	48825.000000	23783.840000	299.026434	24628.000000

	Poblacion	Tasa_Robos	Pobl_Esp_Total	Renta_Media_Alquiler_Em2	Precio_Vivienda_Segunda_Mano_Em2	Renta_Bruta_Disponible
count	1008.000000	1008.000000	1008.000000	1008.000000	1008.000000	1008.000000
mean	152403.833333	48.160714	133152.517857	12.108433	2679.082360	263929.189809
std	53784.456422	49.286770	46424.823051	3.306793	1231.103482	88299.663641
min	45870.000000	0.000000	41831.000000	0.000000	26.000000	0.000000
25%	118734.000000	18.000000	105478.500000	10.113701	1761.000000	186134.242634
50%	143934.500000	36.000000	127082.000000	11.655110	2597.500000	271319.416497
75%	183990.500000	58.000000	166958.250000	14.443200	3521.500000	320171.088801
max	253433.000000	369.000000	225786.000000	19.463076	6043.000000	456397.981257

Fuente: Elaboración Propia

Otro aspecto fundamental es la correlación entre las variables ya que de ello depende el grado de relación lineal (positiva, negativa o nula) existente entre las mismas. Que dos variables estén altamente correladas positivas o negativamente (valores superiores a 0.80) implica que ambas crecen o decrecen de forma paralela y, por consiguiente, da lugar a un patrón de dependencia entre variables. La independencia entre variables lo marca el valor del coeficiente en el valor cero.

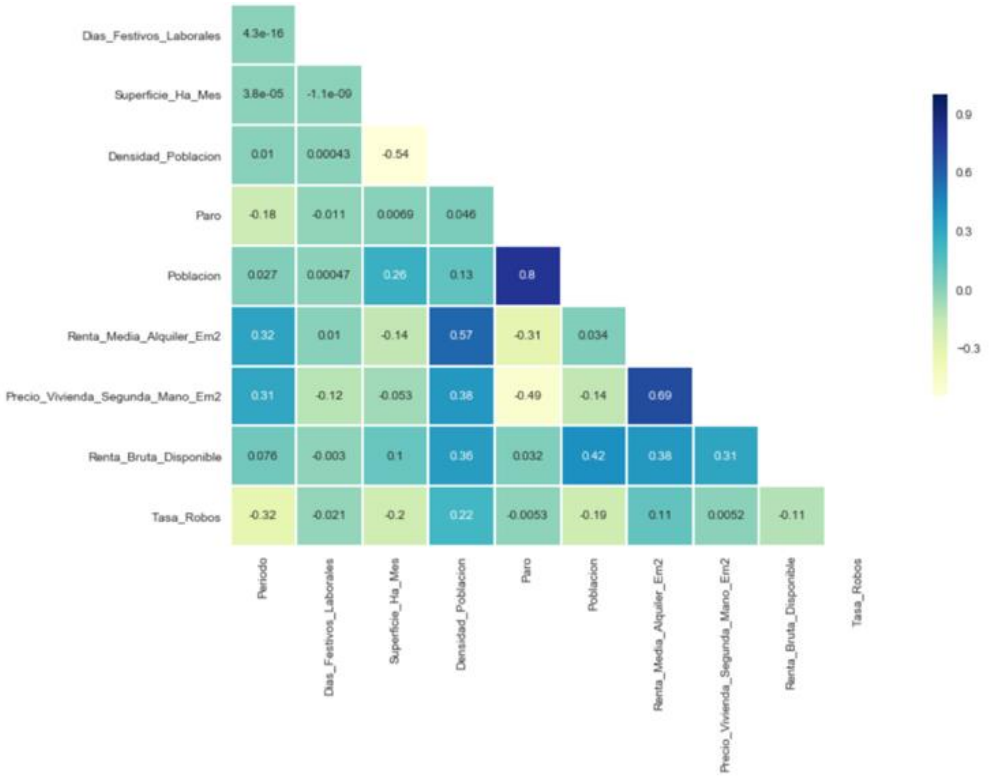


Gráfico 11. Matriz de Correlaciones. Elaboración Propia

Las variables que presentan mayor valor en el coeficiente de correlación son población con paro con un valor de 0,8, seguido del precio de la vivienda con precio del alquiler presentando un valor de 0,69 con renta del alquiler y, por último, renta media alquiler con densidad de población cuyo valor asciende a 0,57.

Por otro lado, es interesante comprobar de forma gráfica cómo se comportan las variables, siendo un caso habitual la utilización del tradicional histograma o gráfico de dispersión que permite comprobar dichos patrones. Sin embargo, en el presente trabajo este análisis se realiza mediante gráficos dinámicos cuyo objetivo se basa en diferenciar cada distrito de Madrid y comprobar las características que presentan cada uno de ellos en función de la variable de interés seleccionada. Gracias a la georreferenciación es posible visualizar la realidad de los datos en las zonas de interés deseadas.

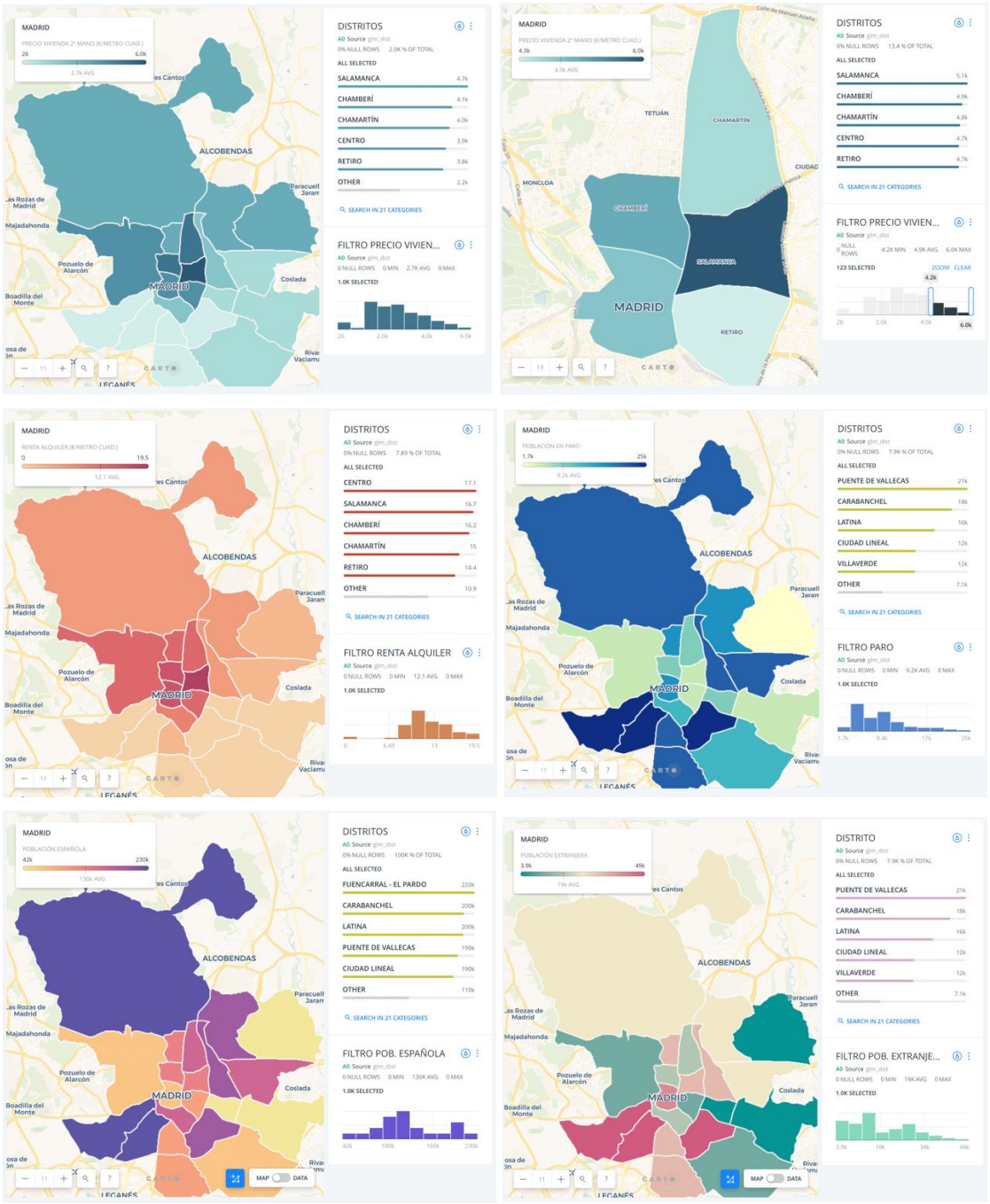


Ilustración 21. Análisis Georreferenciado Variables Cuantitativas por Distritos. Elaboración Propia

Los resultados que se obtienen según la georreferenciación aplicada por variables son los siguientes:

- 🗺 <u>Precio Compra/Venta de Vivienda.</u> El mayor valor asociado al precio de la vivienda corresponde al distrito Salamanca, seguido de los distritos colindantes como son Chamberí, Chamartín, Centro, Retiro y Moncloa. Sin embargo, los distritos Latina, Carabanchel, Usera, Villaverde y Puente de Vallecas son las zonas menos cotizadas.

- Renta de Alquiler. La zona donde se paga un mayor precio por la renta de alquiler corresponde nuevamente a la zona de Salamanca, seguido del distrito de Centro y Chamberí. Por el contrario, la zona sur de Madrid es donde los alquileres presentan unos precios más bajos.
- Paro. Los distritos que mayor población parada presentan son Latina, Carabanchel y Puente de Vallecas, mientras que el distrito de Barajas es el que menor paro se registra, seguido de Salamanca, Retiro y Moncloa.
- Población Española. Los distritos donde se registra un mayor número residentes españoles son Barajas, Moratalaz y Vicálvaro. Mientras, que la zona de menor influencia española pertenece a los distritos de Latina, Carabanchel y Fuencarral - El Pardo.
- Población Extranjera. La residencia mayoritariamente extranjera se localiza en los distritos de Latina, Carabanchel y Puente de Vallecas. Sin embargo, la población extranjera residente en Madrid se reparte en menor medida por los distritos Barajas, Moratalaz y Vicálvaro.

Por último, se lleva a cabo un análisis individual y un poco más exhaustivo para el caso de la frecuencia de robo dado que es la variable que posteriormente se utilizará como respuesta en el modelo de regresión.

Primeramente, se procede a realizar un gráfico de caja y bigotes que permite comprobar de forma visual en función de los diferentes meses los valores que toma dicha variable.

Además, permite evidenciar la existencia de posibles valores extremos o outliers que pueden dar lugar a desviaciones en el posterior modelo. En este caso no se procede a la explicación de los resultados obtenidos ya que al tratarse de un gráfico que es de uso bastante común no resulta de complejidad para el lector poder entenderlo sin problemas.

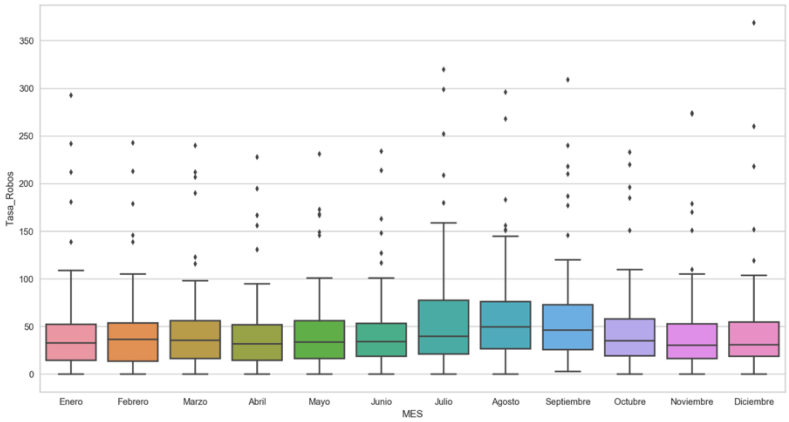


Gráfico 12. Boxplot Variable Respuesta por Mes. Elaboración propia

En segundo lugar, se utiliza el método ***k-means*** que es un método de agrupación de observaciones que genera clústers o grupos basado en distancias, el cual, no se tratará de forma teórica en el presente trabajo ya que es un análisis complementario a nivel exploratorio para el caso de la variable tasa de robos en viviendas.

El objetivo es obtener grupos de observaciones que presentan características homogéneas en función de una primera dimensión en la que se utiliza la variable población en paro y, una segunda dimensión en la que se utiliza la tasa de variable robos en viviendas.

Para seleccionar el número óptimo de clústers se realiza un gráfico de sedimentación.

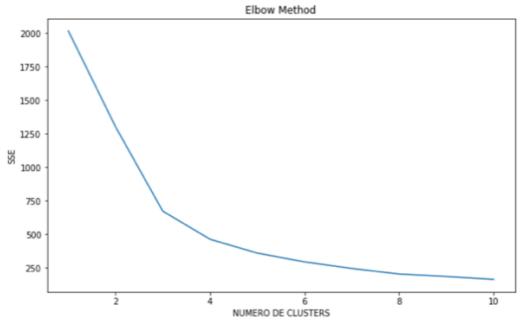


Gráfico 13. Gráfico de Sedimentación (Método Elbow). Elaboración Propia

El gráfico de sedimentación proporciona que el número óptimo de clústers se sitúa entre 2 y 3 grupos. Por esta razón, para obviar complejidades en la interpretación de los resultados se lleva a cabo el método de k-means con una agrupación de k = 2 y k = 3.

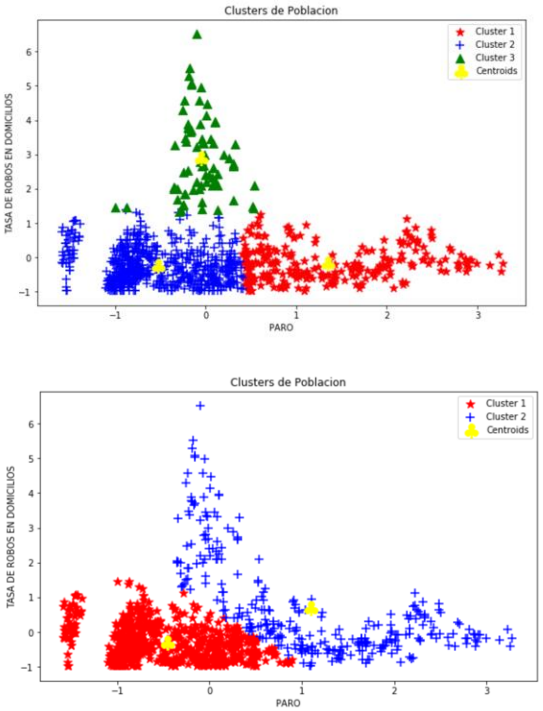


Gráfico 14. Clusters con k = 3 y k = 2. Elaboración Propia

Los resultados que se obtienen en el gráfico para el caso de k = 3 son:

- El *grupo 1* se caracteriza por ser un grupo de personas con un gran índice de criminalidad en robos de viviendas, se podrían considerar tiranos (ladrones).
- El *grupo 2* se caracteriza por una ser un conjunto de población con poco paro y un bajo índice en el número de robos, se pueden catalogar como personas de clase trabajadoras civilizadas que cumplen con las normas de la comunidad.
- El *grupo 3* se caracteriza por ser un conjunto de población con mucho paro y un índice de robos bajo-medio, pudiendo corresponder con personas de clase baja o humilde.

Para completar este apartado se procede a realizar un gráfico georreferenciado, pero en este caso para la frecuencia de robos cogiendo como referencia las zonas donde existe una mayor influencia de criminalidad por la frecuencia de robos en viviendas y se georreferencia la dirección de la Universidad Carlos III pudiendo contrastar la existencia de influencia de criminalidad por este tipo de delito según su ubicación.

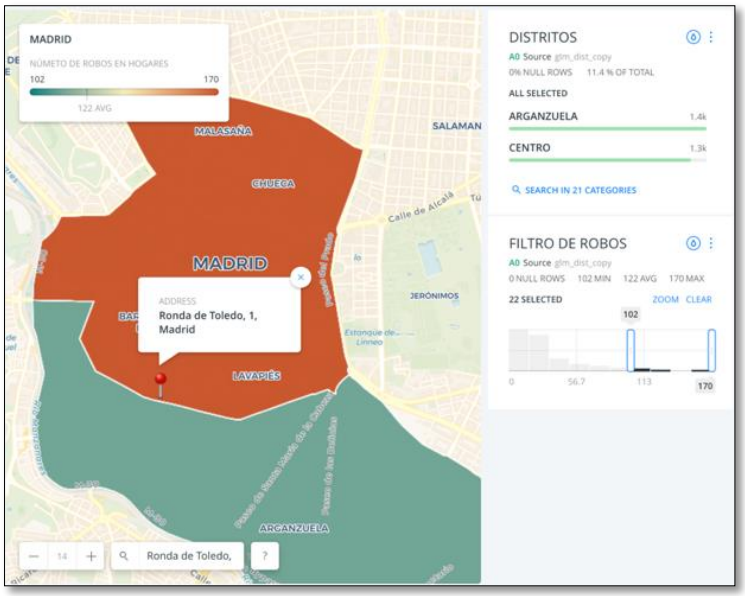


Ilustración 22. Georreferenciación Universidad Carlos III. Elaboración Propia

La Universidad Carlos III se localiza entre los distritos Centro y Arganzuela. Según los resultados se observa que la zona donde se encuentra ubicada presenta un potencial bastante elevado para que se produzca el acaecimiento del siniestro asociado al riesgo de criminalidad por robo en las instalaciones de la misma.

5.3. IMPLEMENTACIÓN DE MODELOS

La elección de modelos que se puede considerar adecuados para su posterior diagnosis no es tarea fácil. La metodología que se ha llevado a cabo para seleccionar tres modelos entre los distintos elaborados ha sido a través del método *Backward* (eliminación hacia atrás), es decir, partiendo del modelo inicial que incluye todas las variables, se eliminan una por una aquellas que aportan menor significación a la variable respuesta hasta conseguir los modelos que se consideran adecuados para el estudio de interés.

La modelización de la tasa de frecuencia de robos en viviendas se ha implementado por medio de la metodología de *Linear Regression Models* (*GLM*) mediante la distribución de Poisson, siendo por excelencia la distribución que se utiliza para modelizar datos de recuento en las actuales compañías.

5.3.1. MODELO I

En primer lugar, se lleva ha cabo la estimación de los parámetros para el *Modelo I*. Para ello, ha sido necesario utilizar como base de código el lenguaje de programación de Python mediante el paquete "*statsmodels*", siendo un software de uso libre utilizado en multitud de disciplinas relacionadas con el tratamiento de datos, aportando un gran potencial en los resultados obtenidos en los distintos análisis.

Para poder llevar a cabo este procedimiento, primeramente, hay que tener en cuenta la codificación de las variables dummies ya que Python no permite su codificación de forma automática a diferencia de otros softwares como es el caso de R, siendo necesario para la posterior validación del modelo. Para ello, se utiliza el paquete "*dummies*" que hace una conversión de las variables categóricas a niveles. Por otro lado, se ha realizado el logaritmo para el caso de las variables que se presentan en unidades monetarias como son el caso de la renta de alquiler y el precio de la vivienda de segunda mano.

Además, al modelizar con una distribución de Poisson la función de enlace (***link***) que selecciona por defecto es la logarítmica. Para el cálculo de la estimación de los parámetros del modelo utiliza el método de mínimos cuadrados ponderados iterativamente (IRLS), siendo un procedimiento que se utiliza para encontrar las estimaciones de máxima verosimilitud para el caso de un GLM. Por último, ha sido necesario introducir un ***offset*** de la variable población de tipo logarítmico como la función enlace, esto permite que se

pondere por la variable que se ha determinado y, en consecuencia, no considerar todas las zonas constantes.

Las variables regresoras que se han tenido en cuenta para la elaboración del primer modelo son de tipo sociodemográficas, económicas, geográficas y temporales, considerándose variables exógenas al riesgo.

Tabla 13. Estimación de los Parámetros Modelo I

Results: Generalized linear model

Model:	GLM	AIC:	16533.0983
Link Function:	log	BIC:	4522.0542
Dependent Variable:	Tasa_Robos	Log-Likelihood:	-8229.5
Date:		LL-Null:	-21818.
No. Observations:	1008	Deviance:	11237.
Df Model:	36	Pearson chi2:	1.03e+04
Df Residuals:	971	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-6.0508	0.0578	-104.6060	0.0000	-6.1642	-5.9375
C(DISTRITOS)[T.Barajas]	2.0006	0.0480	41.6937	0.0000	1.9066	2.0947
C(DISTRITOS)[T.Carabanchel]	-4.3132	0.0644	-66.9516	0.0000	-4.4395	-4.1869
C(DISTRITOS)[T.Centro]	0.6052	0.0214	28.2728	0.0000	0.5633	0.6472
C(DISTRITOS)[T.u'Chamart\xedn']	-0.1513	0.0310	-4.8785	0.0000	-0.2120	-0.0905
C(DISTRITOS)[T.u'Chamber\xed']	-0.7638	0.0366	-20.8935	0.0000	-0.8355	-0.6922
C(DISTRITOS)[T.Ciudad Lineal]	-3.2281	0.0422	-76.4884	0.0000	-3.3108	-3.1454
C(DISTRITOS)[T.Fuencarral-El Pardo]	-3.5042	0.0486	-72.0413	0.0000	-3.5995	-3.4088
C(DISTRITOS)[T.Hortaleza]	-2.3367	0.0381	-61.2764	0.0000	-2.4115	-2.2620
C(DISTRITOS)[T.Latina]	-4.1073	0.0533	-77.1284	0.0000	-4.2117	-4.0030
C(DISTRITOS)[T.Moncloa-Aravaca]	0.2967	0.0329	9.0273	0.0000	0.2323	0.3611
C(DISTRITOS)[T.Moratalaz]	-0.4235	0.0374	-11.3300	0.0000	-0.4968	-0.3503
C(DISTRITOS)[T.Puente de Vallecas]	-4.9317	0.0788	-62.6049	0.0000	-5.0861	-4.7773
C(DISTRITOS)[T.Retiro]	-0.4602	0.0393	-11.6968	0.0000	-0.5373	-0.3831
C(DISTRITOS)[T.Salamanca]	0.2247	0.0274	8.2118	0.0000	0.1711	0.2784
C(DISTRITOS)[T.San Blas-Canillejas]	-1.8530	0.0298	-62.2250	0.0000	-1.9113	-1.7946
C(DISTRITOS)[T.u'Tetu\xeln']	-1.5597	0.0293	-53.1958	0.0000	-1.6172	-1.5022
C(DISTRITOS)[T.Usera]	-1.3011	0.0261	-49.8915	0.0000	-1.3523	-1.2500
C(DISTRITOS)[T.u'Vic\xe1lvaro']	0.2756	0.0487	5.6639	0.0000	0.1802	0.3710
C(DISTRITOS)[T.Villa de Vallecas]	-0.6742	0.0291	-23.1462	0.0000	-0.7313	-0.6171
C(DISTRITOS)[T.Villaverde]	-2.0276	0.0323	-62.7500	0.0000	-2.0909	-1.9643
C(MES)[T.Agosto]	0.2955	0.0131	22.4873	0.0000	0.2697	0.3212
C(MES)[T.Diciembre]	-2.9999	0.0314	-95.6019	0.0000	-3.0614	-2.9384
C(MES)[T.Enero]	-0.0320	0.0271	-1.1782	0.2387	-0.0852	0.0212
C(MES)[T.Febrero]	0.1677	0.0296	5.6638	0.0000	0.1097	0.2258
C(MES)[T.Julio]	0.6291	0.0275	22.8356	0.0000	0.5751	0.6831
C(MES)[T.Junio]	0.3446	0.0284	12.1127	0.0000	0.2888	0.4003
C(MES)[T.Marzo]	-0.0782	0.0254	-3.0773	0.0021	-0.1280	-0.0284
C(MES)[T.Mayo]	0.0093	0.0291	0.3197	0.7492	-0.0478	0.0664
C(MES)[T.Noviembre]	0.1344	0.0279	4.8160	0.0000	0.0797	0.1891
C(MES)[T.Octubre]	0.2134	0.0278	7.6805	0.0000	0.1589	0.2679
C(MES)[T.Septiembre]	0.6806	0.0287	23.7499	0.0000	0.6244	0.7368
C(Estacional_2)[T.No Navidad]	-3.0509	0.0338	-90.1692	0.0000	-3.1173	-2.9846
C(Estacional_3)[T.Vacaciones]	0.2955	0.0131	22.4873	0.0000	0.2697	0.3212
C(Estacional_1)[T.Semana Santa]	0.0917	0.0266	3.4473	0.0006	0.0396	0.1438
Dias_Festivos_Laborales	0.1406	0.0098	14.3938	0.0000	0.1215	0.1598
Paro	0.0003	0.0000	41.6687	0.0000	0.0002	0.0003
Renta_Media_Alquiler_Em2	-0.0232	0.0032	-7.2833	0.0000	-0.0295	-0.0170
Precio_Vivienda_Segunda_Mano_Em2	-0.0001	0.0000	-15.0132	0.0000	-0.0001	-0.0001

Fuente: Elaboración Propia

Estadísticos Individuales

La salida mostrada aporta cada una de las variables regresoras junto con los niveles en caso de tratarse de una variable dummy. Para este tipo de variables el modelo coge como referencia el primer nivel evaluado por orden alfabético, de tal forma, que muestra la

comparativa de los demás niveles en base a la variable de referencia. Las referencias asociadas a cada una de las variables categóricas para el *Modelo I* y el resto de modelos en los que se introduzcan dichas variables son:

- Distrito: Arganzuela
- Mes: Abril
- Estacional 1: No semana Santa
- Estacional 2: Navidad
- Estacional 3: No Vacaciones

La estimación asociada a cada variable lo determinan el valor del coeficiente de regresión que la salida muestra. Este valor indica la variación promedio de la variable respuesta en función de signo del coeficiente del regresor manteniendo constante el resto de variables. De tal forma, que en caso de que el coeficiente sea positivo indicará que manteniéndose el resto de variables constantes la variable respuesta en promedio aumenta e^{β_j} si X_i aumenta en una unidad y, en caso de obtenerse un valor negativo, disminuye en la misma proporción. Esta interpretación cambia en el caso de que los regresores se encuentren en forma de logaritmo, como es el caso de las variables Renta Media de Alquiler y Precio de la Vivienda. Son semielasticidad y su interpretación sería en términos porcentuales, es decir, si X_i aumenta/disminuye en 1% la variable dependiente aumenta/disminuye en $e^{\beta_j/100}$. Para el caso de las variables dummies la interpretación se mantiene a excepción de que la comparativa se realiza en base a la variable de referencia.

Además, la salida proporciona los valores de los p-valores asociados a los estadísticos individuales junto con el intercept o constante del modelo. Las hipótesis de contraste son:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

En el caso de que el p-valor sea menor a un nivel de significación de $\alpha = 0.05$ indica que existe evidencia para rechazar la hipótesis nula y, en consecuencia, la variable regresora es distinta a cero aportando información significativa a la tasa de frecuencia de robos en viviendas. En caso contrario, la variable no aporta relevancia al modelo.

De esta forma, se observa en cómputo general todas las variables junto con el intercept aportan significación a la variable dependiente, a excepción de *enero* y *mayo* que pertenece a dos de los niveles de la variable mes.

Por último, se puede comprobar los valores del error y los intervalos de confianza asociados a cada variable.

Medidas de Bondad de Ajuste

La evaluación del modelo es la parte más importante del análisis, ya que de ello depende de que el modelo seleccionado esté bien especificado para que aporte resultados coherentes y, a su vez, válidos.

En primer lugar, se lleva a cabo una exploración de los residuos que permite comprobar si existe una tendencia lineal en el modelo, además de poder comprobar la hipótesis de normalidad del mismo.

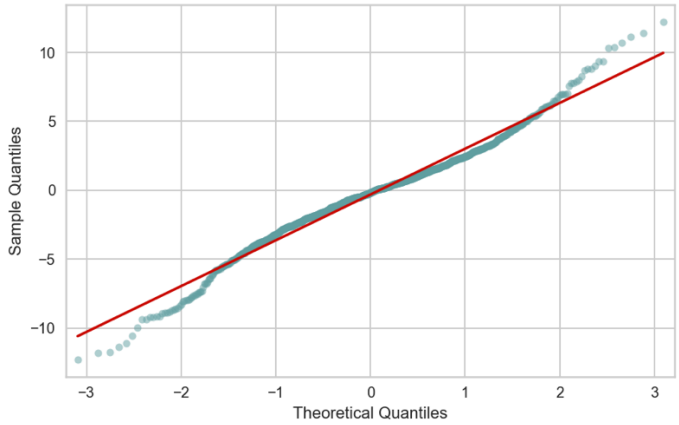


Gráfico 15. Residuos de la Deviance del Modelo I. Elaboración Propia

La evaluación de los residuos confirma que el modelo aparentemente solamente presenta claras deficiencias en los valores superiores e inferiores, pero que en cómputo general parece tener un comportamiento normal.

Para contrastar de formar numérica los resultados. Se lleva a cabo el test de *Jarque-Bera* que permite contrastar la hipótesis de normalidad, siendo las hipótesis de contrastes:

$$\begin{cases} H_0: \exists \text{ Normalidad} \\ H_1: \nexists \text{ Normalidad} \end{cases}$$

El valor del estadístico es 200,64 y el p-valor asociado al mismo de 2.70e-44 considerándose un valor muy próximo a cero y siendo menor a un nivel de significación de $\alpha = 0{,}05$, por lo que existe evidencia para rechazar la hipótesis nula y concluyendo la no existencia de normalidad de los residuos del modelo I.

La falta de normalidad en los residuos del modelo no es un problema para el caso de un modelo lineal generalizado ya que este tipo de modelos permite mayor flexibilidad y no

requiere que se cumpla dicha hipótesis, a diferencia de los modelos lineales clásicos que son más restrictivos.

Para obtener una medida de la precisión del modelo, se lleva a cabo una evaluación de la predicción del mismo, donde se divide el conjunto de datos modelizado en dos subconjuntos, uno de entrenamiento (*train*) y otro de testeo (*test*). En este caso, el modelo se entrena (*train*) con el 80% y, el testeo (*test*) se realiza con el 20 % restante.

A continuación, se realiza la comparativa de las predicciones con los valores reales extraídos para el testeo con el fin de comprobar si el modelo presenta un buen comportamiento.

	Diferencia %	Y_pred	Y_test
283	0.286536	33.094557	33
153	0.345215	21.924053	22
705	0.381949	46.175697	46
836	0.539025	55.698146	56
659	1.182452	62.733120	62
425	1.317762	40.459717	41
418	1.540188	44.306915	45
724	1.572827	47.739229	47
644	1.675725	41.687047	41
608	1.798438	60.061078	59

Tabla 14. Cabecera Errores Porcentuales Modelo I. Elaboración Propia

Por último, se lleva a cabo la curva de predicción de *lasso* que permite comparar la mejor estimación del modelo con las predicciones obtenidas. Esta técnica trata de minimizar la log-verosimilitud negativa, penalizando a los coeficientes de regresión del modelo.

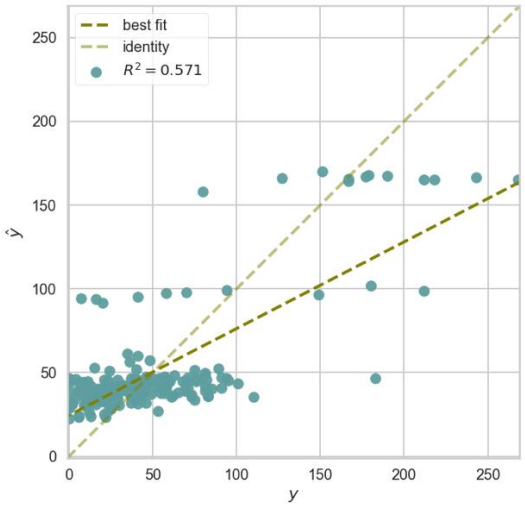


Gráfico 16. Error de Predicción Lasso Modelo I. Elaboración Propia

El modelo presenta una variabilidad del 57,1% considerándose un valor adecuado. Se puede apreciar una deficiencia en la parte superior de la línea horizontal donde se observa que existen algunas predicciones que se sitúan de forma dispersa al resto.

Medidas de Comparación de Modelos

Otras medidas que permiten comprobar la bondad de ajuste en el modelo son las adjuntadas en la *Tabla 15*.

Tabla 15. Medidas Bondad de Ajuste Modelo I

	DEVIANCE	R^2	BIC	AIC	RMSE
Modelo I	11237	57,1%	4522,05	16533,10	372,13

Fuente: Elaboración Propia

Además, son criterios que permiten ser utilizados en la comparación entre modelos pudiendo seleccionar el que mejor características presenten en base a los valores obtenidos, cuyos resultados se detallarán de forma más extensa en el apartado de *Selección de Modelos* para el caso de los tres modelos seleccionados.

5.3.2. MODELO II

La elección del *Modelo II* entre los distintos modelos preliminares realizados es debido a la consistencia de los resultados aportados por las distintas variables regresoras que lo forman.

Las condiciones que se han estipulado para la elaboración del *Modelo II* no difieren de las explicadas anteriormente.

Las variables explicativas que se han tenido en cuenta para la elaboración del Modelo II son variables sociodemográficas, económicas, geográficas y temporales. A diferencia del modelo I en el *Modelo II* se ha introducido la densidad de población como variable explicativa significativa de la tasa de robos en viviendas.

Tabla 16. Estimación de los Parámetros Modelo II

Results: Generalized linear model

Model:	GLM	AIC:	16336.5557
Link Function:	log	BIC:	4330.4273
Dependent Variable:	Tasa Robos	Log-Likelihood:	-8130.3
Date:		LL-Null:	-21818.
No. Observations:	1008	Deviance:	11039.
Df Model:	37	Pearson chi2:	1.01e+04
Df Residuals:	970	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	4.8331	0.7804	6.1930	0.0000	3.3035	6.3627
C(DISTRITOS)[T.Barajas]	-13.4740	1.1077	-12.1643	0.0000	-15.6449	-11.3030
C(DISTRITOS)[T.Carabanchel]	-7.7749	0.2568	-30.2780	0.0000	-8.2782	-7.2716
C(DISTRITOS)[T.Centro]	1.8180	0.0892	20.3743	0.0000	1.6432	1.9929
C(DISTRITOS)[T.u'Chamart\xedn']	-5.6496	0.3944	-14.3252	0.0000	-6.4226	-4.8766
C(DISTRITOS)[T.u'Chamber\xed']	3.0669	0.2766	11.0860	0.0000	2.5247	3.6091
C(DISTRITOS)[T.Ciudad Lineal]	-6.2065	0.2175	-28.5384	0.0000	-6.6327	-5.7802
C(DISTRITOS)[T.Fuencarral-El Pardo]	-18.5360	1.0767	-17.2163	0.0000	-20.6462	-16.4258
C(DISTRITOS)[T.Hortaleza]	-13.7074	0.8144	-16.8316	0.0000	-15.3035	-12.1112
C(DISTRITOS)[T.Latina]	-13.2353	0.6559	-20.1793	0.0000	-14.5208	-11.9498
C(DISTRITOS)[T.Moncloa-Aravaca]	-14.0835	1.0290	-13.6867	0.0000	-16.1003	-12.0667
C(DISTRITOS)[T.Moratalaz]	-5.9659	0.3982	-14.9827	0.0000	-6.7463	-5.1854
C(DISTRITOS)[T.Puente de Vallecas]	-9.6646	0.3486	-27.7263	0.0000	-10.3478	-8.9814
C(DISTRITOS)[T.Retiro]	-1.8969	0.1098	-17.2767	0.0000	-2.1121	-1.6817
C(DISTRITOS)[T.Salamanca]	2.1805	0.1426	15.2896	0.0000	1.9010	2.4601
C(DISTRITOS)[T.San Blas-Canillejas]	-12.8841	0.7898	-16.3123	0.0000	-14.4321	-11.3360
C(DISTRITOS)[T.u'Tetu\xeln']	1.9666	0.2531	7.7693	0.0000	1.4705	2.4627
C(DISTRITOS)[T.Usera]	-5.2551	0.2845	-18.4696	0.0000	-5.8128	-4.6975
C(DISTRITOS)[T.u'Vic\xellvaro']	-14.4154	1.0519	-13.7045	0.0000	-16.4770	-12.3537
C(DISTRITOS)[T.Villa de Vallecas]	-15.1704	1.0374	-14.6232	0.0000	-17.2037	-13.1371
C(DISTRITOS)[T.Villaverde]	-12.8096	0.7723	-16.5855	0.0000	-14.3234	-11.2959
C(MES)[T.Agosto]	0.2903	0.0132	21.9991	0.0000	0.2644	0.3161
C(MES)[T.Diciembre]	2.4460	0.3907	6.2613	0.0000	1.6803	3.2117
C(MES)[T.Enero]	0.0102	0.0273	0.3724	0.7096	-0.0434	0.0637
C(MES)[T.Febrero]	0.1935	0.0298	6.4946	0.0000	0.1351	0.2520
C(MES)[T.Julio]	0.6015	0.0277	21.6890	0.0000	0.5471	0.6558
C(MES)[T.Junio]	0.3413	0.0286	11.9523	0.0000	0.2853	0.3972
C(MES)[T.Marzo]	-0.0751	0.0253	-2.9707	0.0030	-0.1246	-0.0255
C(MES)[T.Mayo]	0.0052	0.0291	0.1773	0.8593	-0.0518	0.0622
C(MES)[T.Noviembre]	0.1850	0.0281	6.5727	0.0000	0.1298	0.2401
C(MES)[T.Octubre]	0.2569	0.0281	9.1577	0.0000	0.2019	0.3119
C(MES)[T.Septiembre]	0.6841	0.0288	23.7553	0.0000	0.6276	0.7405
C(Estacional_2)[T.No Navidad]	2.3871	0.3903	6.1154	0.0000	1.6220	3.1522
C(Estacional_3)[T.Vacaciones]	0.2903	0.0132	21.9991	0.0000	0.2644	0.3161
C(Estacional_1)[T.Semana Santa]	0.1238	0.0266	4.6550	0.0000	0.0717	0.1759
Dias_Festivos_Laborales	0.1415	0.0098	14.4546	0.0000	0.1223	0.1606
Densidad_Poblacion	-0.0673	0.0048	-13.9784	0.0000	-0.0768	-0.0579
Paro	0.0002	0.0000	25.5163	0.0000	0.0002	0.0002
Renta_Media_Alquiler_Em2	-0.0227	0.0032	-7.1291	0.0000	-0.0289	-0.0164
Precio_Vivienda_Segunda_Mano_Em2	-0.0001	0.0000	-13.3178	0.0000	-0.0001	-0.0001

Fuente: Elaboración Propia

Estadísticos Individuales

Las hipótesis de contrastes para los coeficientes de regresión individuales para el modelo II son las siguientes:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

De igual forma que el caso anterior, se observa que a excepción de la variable *enero* y *mayo* el resto de variables junto con el intercept presentan unos p-valores iguales al valor

cero rechazando la hipótesis nula a un nivel de confianza del 95% y, en consecuencia, aportando información significativa a la tasa de robos en viviendas.

Medidas de Bondad de Ajuste

Para contrastar si el modelo está bien especificado se realiza un análisis de los residuos del mismo tanto de forma gráfica como numérica.

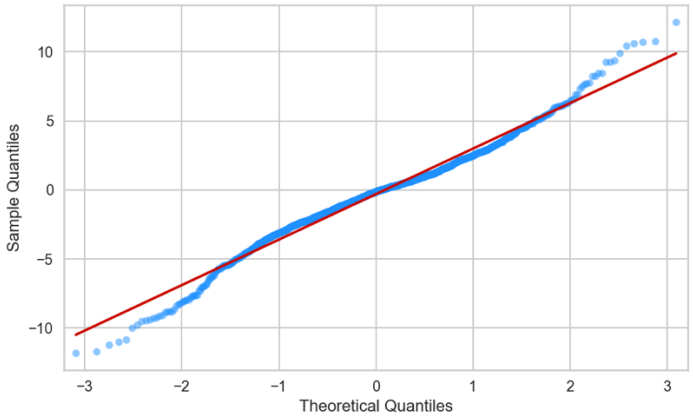


Gráfico 17. Residuos de la Deviance del Modelo II. Elaboración Propia

El *Modelo II* presentan desviaciones en los residuos. Visualizando el gráfico parece que tener un comportamiento normal a excepción de los extremos donde los desvíos son más pronunciados.

Para contrastar de forma numérica los resultados obtenidos en el gráfico se lleva a cabo el valor del test de normalidad asociado a los residuos del *Modelo II*.

$$\begin{cases} H_0: \exists \text{ Normalidad} \\ H_1: \nexists \text{ Normalidad} \end{cases}$$

El test de *Jaque-Bera* presenta un valor del estadístico de contraste de 169.30 y un p-valor igual a 1.72e-37 siendo un valor muy próximo a cero y menor a un nivel de significación de $\alpha = 0{,}05$ por lo que se asume la no existencia de normalidad para los residuos en el modelo II.

Para comprobar el comportamiento del *Modelo II* se realiza un entrenamiento (*train*) del modelo y, posteriormente, se testea (*test*) cuyos resultados se ajunta en el *Gráfico 18*. Además, se muestra la salida de los diez primeros valores del error porcentual calculado a partir de la diferencia entre los valores predichos y observados del modelo.

	Diferencia %	Y_pred	Y_test
608	0.003555	58.997903	59
659	0.386181	61.760568	62
513	0.417886	33.857919	34
836	0.553982	55.689770	56
67	0.643050	6.038583	6
418	0.694846	44.687319	45
639	0.721544	44.317479	44
345	1.101674	18.198301	18
555	1.779551	33.394953	34
458	2.105059	41.115875	42

Tabla 17. Cabecera Errores Porcentuales Modelo II. Elaboración Propia

El gráfico de la curva de predicción de Lasso muestra por un lado la identidad que es equivalente a la curva horizontal y, por otro lado, la mejor estimación que el modelo podría ofrecer. Los puntos son las predicciones asociadas a nuestro *Modelo II*, con lo cual cuanto más cercano estén de la curva de mejor estimación mejor comportamiento tendrá el mismo.

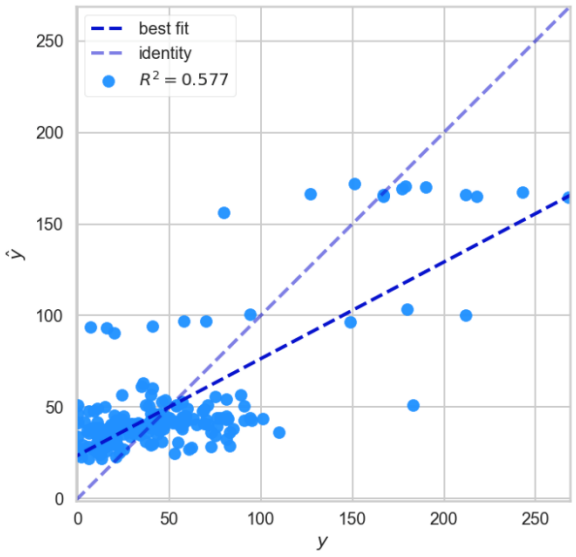


Gráfico 18. Error de Predicción Lasso Modelo II. Elaboración Propia

El modelo II presenta una variabilidad del 57,7% considerándose un valor moderado. Se puede apreciar una agrupación de los valores en la parte inferior de la curva y algunos puntos dispersos. Los puntos que divagan de forma individualizada podrían considerarse posibles valores atípicos u outliers pero esta confirmación no es decisiva ya que para ello sería necesario realizar un análisis más exhaustivo.

Medidas de Comparación de Modelos

Los valores asociados a las medidas de bondad de ajuste para definir si el modelo es adecuado se adjunta en la *Tabla 5*. Estos valores se utilizan en el siguiente apartado para contrastar la comparación entre los distintos modelos establecidos ya que de forma individualizada no aportan demasiada información, a excepción del coeficiente de determinación que aporta la variabilidad de la tasa de robos de viviendas que queda explicada por el *Modelo II*.

Tabla 18. Medidas Bondad de Ajuste Modelo II

	DEVIANCE	R^2	BIC	AIC	RMSE
Modelo II	11039	57,7%	4330,43	16336,56	370,35

Fuente: Elaboración Propia

5.3.3. MODELO III

En la construcción del *Modelo III* se han eliminado las variables temporales 1 y 2, mes y densidad de población

Tabla 19. Estimación de los Parámetros Modelo III

Results: Generalized linear model

Model:	GLM	AIC:	17941.6149
Link Function:	log	BIC:	5871.5822
Dependent Variable:	Tasa Robos	Log-Likelihood:	-8945.8
Date:		LL-Null:	-21818.
No. Observations:	1008	Deviance:	12670.
Df Model:	24	Pearson chi2:	1.18e+04
Df Residuals:	983	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-8.2052	0.0787	-104.2803	0.0000	-8.3594	-8.0509
C(DISTRITOS)[T.Barajas]	1.6711	0.0451	37.0721	0.0000	1.5827	1.7594
C(DISTRITOS)[T.Carabanchel]	-3.8011	0.0596	-63.8082	0.0000	-3.9178	-3.6843
C(DISTRITOS)[T.Centro]	0.6636	0.0210	31.5658	0.0000	0.6224	0.7048
C(DISTRITOS)[T.u'Chamart\xedn']	-0.2812	0.0305	-9.2181	0.0000	-0.3410	-0.2214
C(DISTRITOS)[T.u'Chamber\xed']	-0.8752	0.0363	-24.1372	0.0000	-0.9463	-0.8041
C(DISTRITOS)[T.Ciudad Lineal]	-3.0081	0.0407	-73.8229	0.0000	-3.0879	-2.9282
C(DISTRITOS)[T.Fuencarral-El Pardo]	-3.3929	0.0483	-70.2022	0.0000	-3.4876	-3.2982
C(DISTRITOS)[T.Hortaleza]	-2.3030	0.0381	-60.5051	0.0000	-2.3776	-2.2284
C(DISTRITOS)[T.Latina]	-3.7369	0.0503	-74.3430	0.0000	-3.8355	-3.6384
C(DISTRITOS)[T.Moncloa-Aravaca]	0.1212	0.0319	3.8060	0.0001	0.0588	0.1837
C(DISTRITOS)[T.Moratalaz]	-0.5696	0.0364	-15.6275	0.0000	-0.6410	-0.4982
C(DISTRITOS)[T.Puente de Vallecas]	-4.2870	0.0724	-59.2022	0.0000	-4.4290	-4.1451
C(DISTRITOS)[T.Retiro]	-0.6250	0.0386	-16.1742	0.0000	-0.7007	-0.5493
C(DISTRITOS)[T.Salamanca]	0.1218	0.0270	4.5103	0.0000	0.0689	0.1747
C(DISTRITOS)[T.San Blas-Canillejas]	-1.7693	0.0295	-59.8796	0.0000	-1.8272	-1.7114
C(DISTRITOS)[T.u'Tetu\xeln']	-1.4885	0.0291	-51.1082	0.0000	-1.5455	-1.4314
C(DISTRITOS)[T.Usera]	-1.1984	0.0257	-46.5560	0.0000	-1.2489	-1.1480
C(DISTRITOS)[T.u'Vic\xe1lvaro']	0.0644	0.0474	1.3588	0.1742	-0.0285	0.1573
C(DISTRITOS)[T.Villa de Vallecas]	-0.7196	0.0289	-24.8919	0.0000	-0.7762	-0.6629
C(DISTRITOS)[T.Villaverde]	-1.8434	0.0313	-58.8440	0.0000	-1.9048	-1.7820
C(Estacional_3)[T.Vacaciones]	0.3146	0.0149	21.0492	0.0000	0.2853	0.3439
Paro	0.0002	0.0000	36.4917	0.0000	0.0002	0.0002
Renta_Media_Alquiler_Em2	-0.0268	0.0031	-8.5314	0.0000	-0.0330	-0.0206
Precio_Vivienda_Segunda_Mano_Em2	-0.0001	0.0000	-20.4510	0.0000	-0.0001	-0.0001

Fuente: Elaboración Propia

Estadísticos Individuales

Las hipótesis de contraste son las siguientes:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

Todas las variables regresoras del *Modelo III* presentan un p-valor significativo a un nivel de confianza del 95% ya que dichos valores son nulos indicando que aportan relevancia significativa para explicar la tasa de robos de viviendas, a excepción del distrito de Vicálvaro que acepta la hipótesis nula de no significación estadística para la variable respuesta.

Medidas de Bondad de Ajuste

Los residuos del Modelo III se muestran de forma gráfica pudiendo ser visualizados de una forma sencilla para comprobar si tienden a un comportamiento gaussiano.

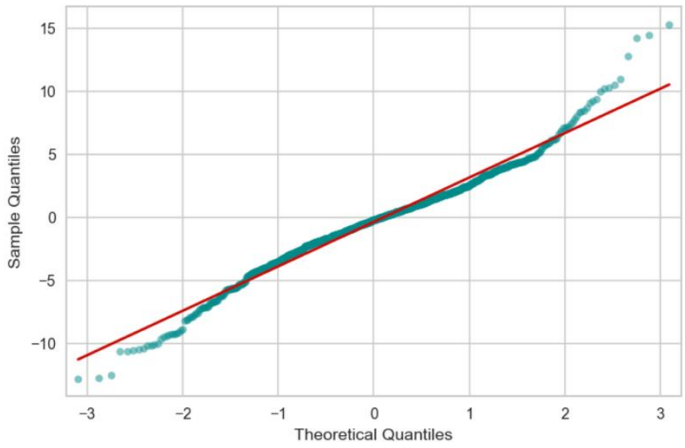


Gráfico 19. Residuos de la Deviance del Modelo III. Elaboración Propia

Los residuos del *Modelo III* presentan una dudosa apariencia de normalidad. Para obtener una respuesta más exacta los resultados del gráfico se contrastan de forma numérica con el test de normalidad para los resiudos.

$$\begin{cases} H_0: \exists \text{ Normalidad} \\ H_1: \nexists \text{ Normalidad} \end{cases}$$

El test de *Jarque-Bera* permite contrastar si los residuos siguen una distribución normal verificando que los valores de asimetría y curtosis se comportan de forma gaussiana. El valor del estadístico es 551,44 y el p-valor asociado de 1.80e-120 siendo menor a un nivel de significación de $\alpha = 0{,}05$ rechazando la hipótesis nula de que los residuos se distribuyen normalmente.

	Diferencia %	Y_pred	Y_test
235	0.047216	21.009915	21
150	0.049650	52.973685	53
679	0.477354	12.057282	12
602	0.505650	24.873588	25
724	0.584641	47.274781	47
916	0.970477	56.446828	57
530	1.041275	44.531426	45
705	1.517816	46.698195	46
552	1.565445	26.407016	26
604	1.579216	56.099847	57

Tabla 20. Cabecera Errores Porcentuales Modelo III. Elaboración Propia

La curva Lasso indica que el *Modelo III* presenta una variabilidad del 57,2%. Se puede visualizar que el modelo presenta una agrupación en la zona inferior, pero con valores que no llegan a superponerse encima de la mejor estimación.

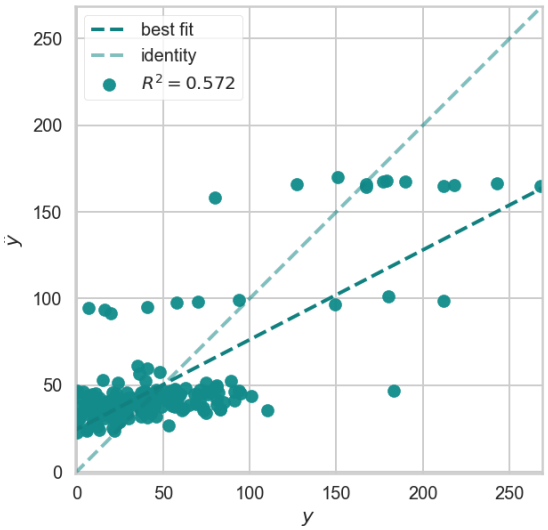


Gráfico 20. Error de Predicción Lasso Modelo III. Elaboración Propia

Medidas de Comparación de Modelos

Estas medidas servirán de métricas comparativas para el apartado de Selección de Modelos.

Tabla 21. Medidas Bondad de Ajuste Modelo III

	DEVIANCE	R^2	BIC	AIC	RMSE
Modelo III	12670	57,2%	5871,58	17941,61	376,43

Fuente: Elaboración Propia

5.3.4. SELECCIÓN DE MODELOS

Tras la fase de especificación y evaluación de los modelos se deduce que los modelos preliminares seleccionados son moderadamente consistentes para explicar la rasa de robos por viviendas.

La *Tabla 22* presenta una comparativa de las principales medidas que se utilizan para la comparación de modelos pudiendo seleccionar aquel que mejor se ajuste a los objetivos preestablecidos en cada caso.

Tabla 22. Medidas Selección de Modelos

	VARIABLES	VARIABLES SIG.	DEVIANCE	R^2	BIC	AIC	RMSE
Modelo I	9	8	11237	57,10%	4522,05	16533,10	372,13
Modelo II	10	9	11039	57,7%	4330,43	16336,56	370,35
Modelo III	5	4	12670	57,2%	5871,58	17941,61	376,43

Fuente: Elaboración Propia

El mejor modelo entre los seleccionados es aquel que presenta un menor valor de la deviance, criterio de información Bayesiano (BIC), criterio de información Akaike (AIC) y error cuadrático medio (RMSE). Además, debe de presentar un valor elevado del coeficiente de regresión puesto que es la variabilidad de la tasa de robos en viviendas que queda explicada por el modelo de regresión. Por último, debe de cumplir el criterio de parsimonia, es decir, los modelos más sencillos son los que mejores se comportan y, en consecuencia, los más deseados.

En base a los requisitos que debe de cumplir un modelo para considerarse competente se selecciona el ***Modelo I***, ya que entre los dos mejores modelos las diferencias son muy reducidas.

Otro motivo de la elección de este modelo es que se obvia introducir la variable densidad de población que a pesar de que los valores de los factores de inflación calculados no ofrecen indicios de multicolinealidad, se podría esta sobreestimando el modelo al considerar una variable que ha sido introducida de igual forma que para el cálculo de la tasa de robo en viviendas como es el caso de la variable población.

Es importante mencionar que la variabilidad de la tasa de robos en viviendas que queda explicada por el Modelo I es del 57,1 % considerándose un valor moderado. Esta

variabilidad es difícil incrementarla dado que en el presente trabajo el tamaño de las observaciones es algo reducido, siendo uno de los puntos a destacar para futuras investigaciones.

5.4. MODELO GEORREFERENCIADO MEDIANTE SCORE

5.4.1. HERRAMIENTA GEOESPACIAL

En la actualidad, existen múltiples programas que permiten trabajar con *SIG*. En el presente trabajo, la georreferenciación de la tasa de robos en hogares se realiza mediante la plataforma de *Location Intelligence* de ***CARTO***, siendo considerada una de las herramientas de posicionamiento geoespacial mejor valoradas del momento a nivel internacional.

Cabe mencionar que esta plataforma no es de uso libre. Sin embargo, contiene una versión de prueba que permite hacer uso limitado de las ventajas que la plataforma ofrece.

¿Por qué CARTO? Principalmente, porque no es solamente una herramienta que permite realizar mapeos de datos con el objetivo de poder interpretarlos, sino que incorpora el valor añadido de poder realizar análisis internos e interactuar con la plataforma mediante los lenguajes de programación ***Python*** y ***PostgreSQL***.

De esta forma, la georreferenciación de la tasa de robos en hogares realizada con ***CARTO*** se compone de dos partes fundamentales:

- *Modelización* de la tasa de robos como variable respuesta mediante el uso de la metodología de modelos lineales generalizados con la distribución de Poisson, cuyo objetivo es obtener un modelo explicativo realizado mediante el lenguaje de ***Python***.
- Localización de aquellas zonas con potencialidad de criminalidad para este tipo de delitos diferenciadas mediante un *Score*. El Score o puntuaciones entre zonas, se realiza teniendo en cuenta el valor de la estimación para la tasa de robos en viviendas ponderadas entre los valores 0 y 1 para los valores de las variables asociadas en cada caso.

Para llevar a cabo la georreferenciación y poder visualizar los resultados obtenidos se toma en consideración la capa geográfica a nivel *distrito*, pertenecientes a la ciudad de

Madrid. Además, de forma complementaria para poder mostrar que se puede realizar intersecciones geoespaciales con la herramienta, se le añade una segunda capa a nivel *barrio*, siendo los valores asociados a cada uno el valor del distrito que corresponda, al no disponer de datos asociados a esta segunda capa.

El mapeo de los datos se realizada de dos formas diferentes, siendo una complementaria a la otra. Por un lado, se realiza mediante el lenguaje de programación ***Python*** pudiendo exponer el uso del mismo para implementar datos georreferenciados.

Por otro lado, con el objetivo de realizar un mapeo interactivo más visual, se utiliza ***CARTO builder,*** que a través de los datos que se han obtenido anteriormente permite crear el mismo mapa con el añadido de poder incorporarle las métricas de interés y, a su vez, incluir filtrados de las distintas zonas o niveles geográficos georreferenciados en función de las mismas.

Como se ha comentado en varias ocasiones a lo largo del presente trabajo, el objetivo principal es construir un modelo explicativo para la tasa de robos en hogares y, poder valorar mediante un *Score* las zonas de mayor potencial en criminalidad. Por ello, en este caso, no solamente se realiza un análisis de las zonas que presentan un mayor o menor *Score* en función de los valores obtenidos, sino también se tiene en cuenta cómo varía la tasa de robos en función de las variables explicativas del modelo pudiendo encontrar diferencias para las distintas zonas. En cualquiera de los casos expuestos, el mapa de calor estará asociado al valor de *Score* que se obtenga para cada zona geográfica.

5.4.1.1. IMPLEMENTACIÓN CON PYTHON

Primeramente, se realiza una comparativa agrupando los distritos de Madrid en dos grupos distintos ya que esto permite poder comprobar si existen diferencias significativas entre la zona norte y sur de la ciudad en función de la menor o mayor potencialidad asociada a la tasa de robos en viviendas.

Ilustración 23. Distritos con Mayor y Menor tasa de robos de Viviendas. Elaboración Propia

Los resultados indican que las zonas más seguras son mayoritariamente la zona norte, mientras que la tasa de robos en la zona sur aumenta respecto a la misma.

Para obtener una mayor información sobre la tasa de robos en viviendas por distrito se divide los niveles en cuatro grupos diferentes mediante el método de agrupación de *Jenks* que se basa en minimizar la desviación promedio en cada grupo, siendo:

- *Grupo 1*: Distritos con potencialidad en robos de viviendas muy alta.
- *Grupo 2*: Distritos con alta potencialidad en robos de viviendas.
- *Grupo 3*: Distritos con moderad potencialidad en robos de viviendas.
- *Grupo 4*: Distritos con potencialidad en robos de viviendas baja.

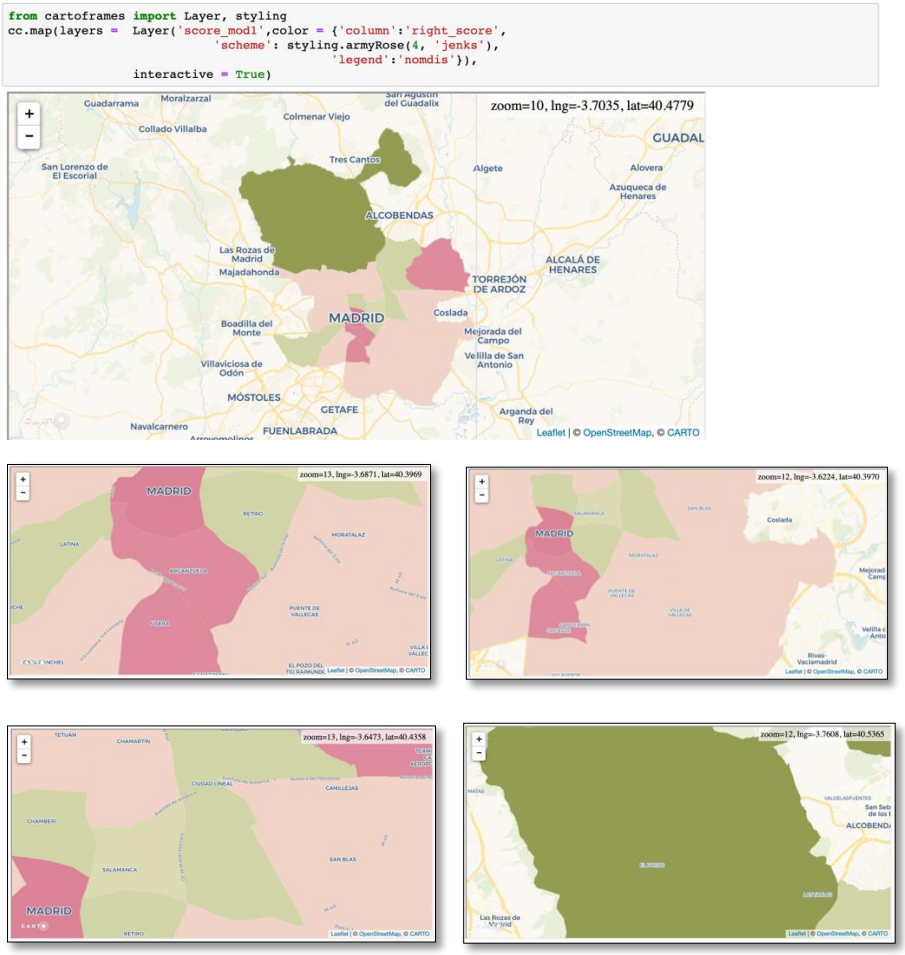


Ilustración 24. Comparativa de las Zonas Potencialidad de Robos en Viviendas. Elaboración Propia

Según esta agrupación se puede observar como los distritos de Arganzuela, Centro, Usera y Barajas son las zonas que presentan una mayor tasa de robos en viviendas, seguido de los distritos que pertenecen principalmente a la zona sur y noroeste de Madrid como son Villaverde, Vallecas, Moratalaz, Carabanchel, San Blas y Vicálvaro al sur, junto con el distrito de Moncloa, Aravaca, Chamartín y Tetuán al noroeste.

Las zonas con un potencial moderado se podrían dividir en dos subgrupos:

- Distritos con alto poder adquisitivo, donde se encuentran los barrios más lujosos de Madrid como son Chamberí, Retiro y Salamanca donde habitualmente se instalan sistemas de seguridad en las viviendas.
- Distritos que presentan una calidad de vida más baja que los anteriores, es decir, son considerados más pobres debido a que presentan una menor renta familiar bruta per cápita. A este subgrupo pertenece los distritos de Hortaleza, Ciudad Lineal y Latina.

Por último, los distritos que presentan una menor tasa de robos en viviendas y, en consecuencia, un Score más bajo son principalmente los situados en la zona norte de Madrid, como es el distrito de Fuencarral – El Pardo.

5.4.1.2. IMPLEMENTACIÓN CON CARTO BUILDER

Las métricas que se van a tener en consideración para representar de forma visual la tasa de robos en viviendas en el builder de CARTO son las variables que han sido introducidas en el *Modelo I* (detallado en la *Tabla 13*) seleccionada y validadas anteriormente, a excepción de las variables temporales al tratarse de variables de carácter cualitativo.

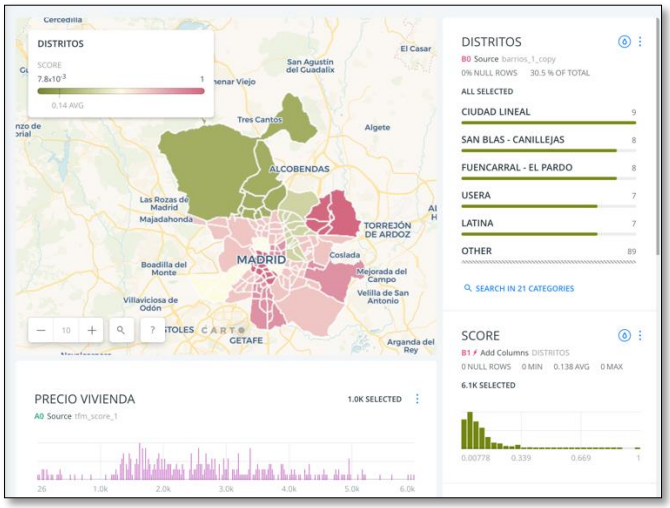


Ilustración 25. Score Tasa de Robos en Viviendas CARTO Builder. Elaboración Propia

Además, para analizar qué variación experimenta la tasa de cambio en función de la métrica seleccionada se realiza un mapeo para cada una de ellas filtrando por el mayor y menor valor correspondiente dichas variables, cuya explicación se adjunta de forma detallada en el apartado de *Conclusiones y futuras vías de desarrollo*.

- Score

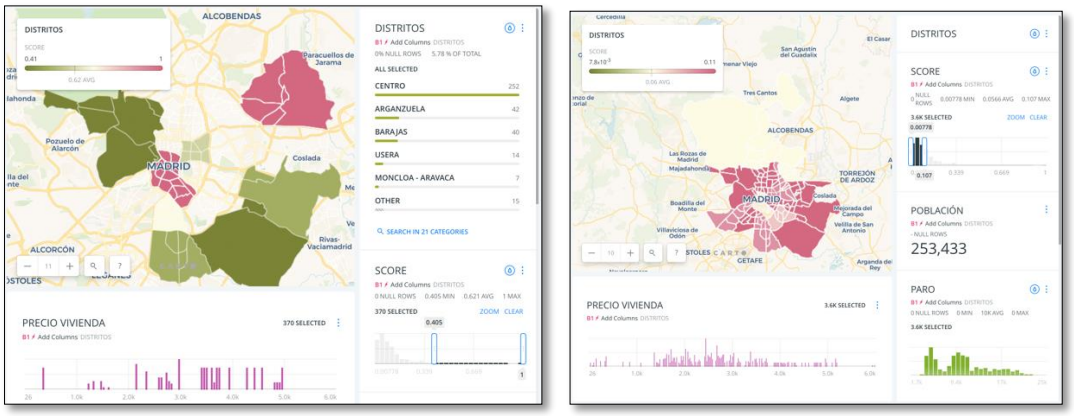


Ilustración 26. Tasa de Robos en Viviendas por Score. Elaboración Propia

- Precio de la vivienda de segunda mano

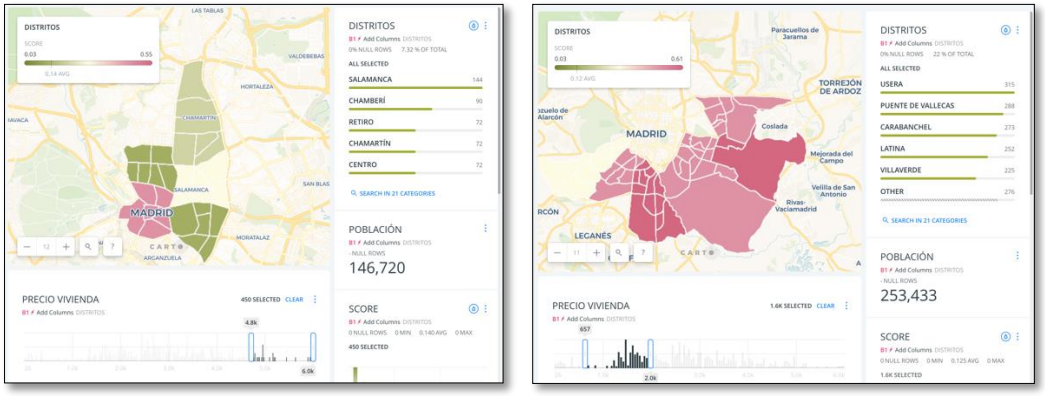


Ilustración 27. Tasa de Robos en Viviendas por Precio. Elaboración Propia

- Paro

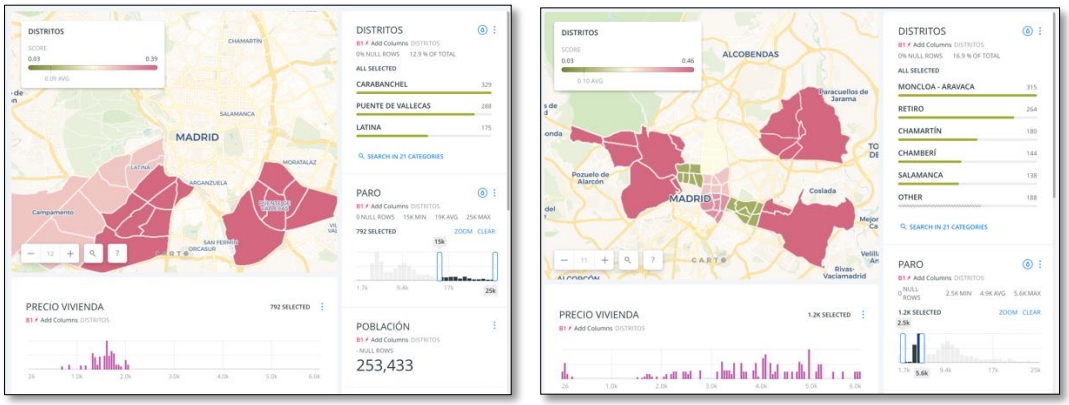


Ilustración 28. Tasa de Robos en Viviendas por Población Parada. Elaboración Propia

- Precio medio del alquiler

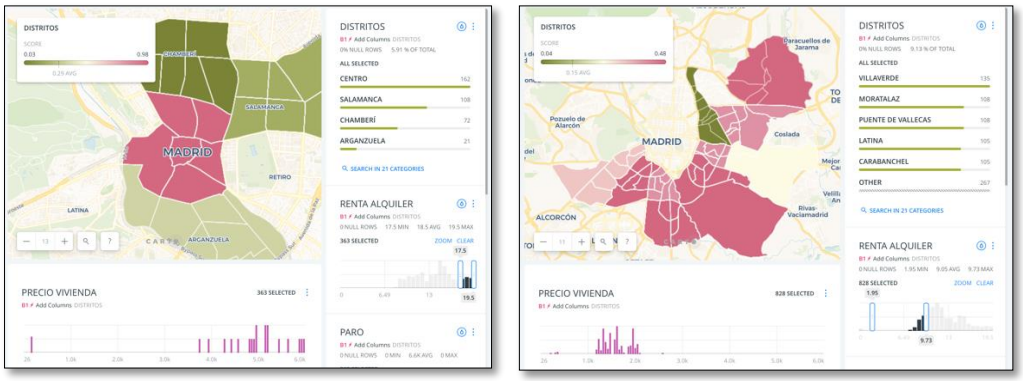


Ilustración 29. Tasa de Robos en Viviendas por Precio Alquiler. Elaboración Propia

6. CONCLUSIONES Y FUTURAS VÍAS DE DESARROLLO

Tras realizar los análisis correspondientes se concluye que el ***Modelo I*** es considerado como un modelo válido para explicar la tasa de robos en viviendas aportando una variabilidad del 57,1%, no pudiendo ser incrementada debido a la escasez en el número de observaciones.

Además, cabe remarcar que la proporción que corresponde a la criminalidad de robos con fuerza en domicilios presenta un valor bajo en todos los años analizados, rondando entre el 1 y 5% respecto al total de delitos. Esto conlleva a que las diferencias en la variación de la tasa de robos en viviendas entre los distintos distritos se reduzcan y presenten valores cercanos.

La variable geográfica ***distritos*** aporta significación estadística al modelo. Al tratarse de una variable dummy y contener distintos niveles, se aprecia que en comparación al nivel de referencia que en este caso es el distrito de *Arganzuela*, la tasa de robos en viviendas varía. De esta forma, en caso de residir en los distritos *Centro*, Barajas, *Moncloa*, *Salamanca* y *Vicálvaro* en lugar del distrito de *Arganzuela* la probabilidad de que invadan la intimidad del hogar aumenta, mientras que para el resto distritos disminuye, respectivamente.

Si se toma en consideración los resultados de la georreferenciación realizada por medio del *Score* de la tasa de robos en viviendas, la zona más tranquila pertenece al distrito de *Fuencarral - El Pardo*, situada en la zona norte de Madrid. Sin embargo, las viviendas ubicadas en las zonas del distrito *Centro*, *Arganzuela* y *Barajas*, seguido de las viviendas situadas en zona sur de la ciudad donde encabezan la lista los distritos de *Usera*, *Villaverde* y *Vicálvaro*, presentan un mayor potencial en la criminalidad de este tipo de delito; siendo la zona más conflictiva el distrito *Centro*, alcanzando el valor máximo del *Score* en la tasa de robos.

La variable ***paro*** presenta un valor positivo, indicando que a medida que la población en situación de desempleo aumenta, la tasa de robo en viviendas también lo hace. Esto se debe a que el desempleo genera mayor nivel de pobreza y se asocia a que un alto desempleo hace incrementar los delitos.

La georreferenciación de la variable ***paro*** permite comprobar que los distritos donde se concentran la mayor proporción de población desempleada pertenecen de nuevo a la zona sur de Madrid, siendo los distritos de *Carabanchel*, *Puente de Vallecas* y *Latina* los más

destacados; presentando una tasa de robos en viviendas que oscila entre 0,03 y 0,39, respectivamente. Sin embargo, las zonas con menor paro corresponden a los distritos que se consideran los más ricos de la ciudad, destacando el distrito de *Salamanca*, *Chamartín*, *Chamberí* y *Retiro*, además de los distritos de *Moncloa*, *Barajas* y *Vicálvaro*, donde estos último lideran como las zonas que presentan mayores valores en el *Score* por criminalidad, oscilando entre 0,03 y 0,46.

El ***número de días no laborales*** al mes también afecta de forma positiva a la tasa de robos en viviendas. El estudio realizado por (UNESPA, 2014) confirma este resultado ya que *"desmonta el mito de que los robos se concentran en los fines de semana"* y confirmando que *"el peor día en materia de robos en la Comunidad de Madrid parece ser el lunes, seguido del jueves".*

Las variables de ***temporalidad*** son también influyentes. En vacaciones de *verano*, *Semana Santa* y *Navidad* la tasa de robos en viviendas aumenta con respecto a otras fechas que no son las señaladas. Esto es debido a que en estas épocas las familias se encuentran fuera de casa y, en consecuencia, la vulnerabilidad de la intimidad del hogar se incrementa.

El ***mes*** es otra de las variables que aportan información significativa a la tasa de robos en viviendas. Para los meses de *diciembre*, *marzo* y *enero* la tasa de robos en viviendas disminuye con respecto al mes de *abril*, mientras que para el resto de los meses aumenta, siendo las épocas veraniegas (*julio*, *agosto* y *septiembre*), las fechas en las cuales se producen mayor tasa de robos en los hogares madrileños.

Por otro lado, la tasa de robos en viviendas disminuye cuando aumenta la renta media del ***alquiler*** y el ***precio*** de la vivienda de segunda mano.

Las zonas caras y lujosas de la ciudad corresponden con toda la zona centro de Madrid, siendo los distritos *Centro*, *Salamanca*, *Chamberí*, *Retiro* y *Chamartín* los distritos destacados por altos ***precios*** de ***compra/venta*** de viviendas, tanto para el caso de primera como de segunda mano. La tasa de robos en estas zonas varía entre 0,03 y 0,98, siendo este último, el valor del ***Score*** que corresponde con el distrito *Centro*, zona donde la criminalidad se considera elevada por alcanzar el valor máximo. La peor zona para invertir es la zona sur de la ciudad, siendo *Carabanchel*, *Usera* y *Villaverde* los distritos menos valorados. El valor del *Score* de la tasa de robos en viviendas asociado a estas zonas se incrementa al valor 0,61.

Respecto al precio de la renta de ***alquiler***, en la zona centro es donde se encuentran las cuantías más elevadas, siendo *Salamanca*, *Chamberí*, *Arganzuela* y *Centro* los distritos más diferenciados. En contraposición, las zonas más económicas son principalmente los distritos que pertenecen a la zona sur, como son *Carabanchel*, *Villaverde*, *Vallecas*, *Moratalaz* y *Carambachel*, además de Barajas presentando un valor del Score de 0,48.

De esta forma, se comprueba que el objetivo de poder ubicar los distritos que presentan mayor exposición al riesgo de robo en viviendas para la ciudad de Madrid ha sido logrado, pudiendo implementarse esta metodología para la fase de tarificación del seguro de hogar para dicha cobertura a los datos de una compañía.

Por último, aportar como consideración en ***futuras líneas de investigació*n** que se deja abierta la posibilidad de poder realizar modelos que contengan un mayor número de variables con información geográfica, como pueden ser puntos de interés que pudieran ser interesante introducir como nuevas variables (cercanía a redes de transporte, centros comerciales, parques de bomberos, comisarías de policías…), no habiendo sido posible su incorporación por limitaciones en la versión de prueba del software utilizado. Utilizar variables geográficas que tengan relevancia en el modelo hace que la variabilidad de este aumente y, el error disminuya, pudiendo obtener resultados más precisos para la toma de decisiones.

Además, sería interesante el estudio e implementación de otros métodos estadísticos tanto con el objetivo de predecir como de explicar, pudiendo utilizarse ***Random Forest***, ***Decisions Tree***, ***Clusters***, entre otros, combinando SIG en función del objetivo de estudio con datos reales de una compañía, a un nivel de capa inferior. En el caso del seguro de hogar y, para la cobertura de robo, el estudio podría ser realizarlo a nivel código postal, siendo una variable que forma parte de cualquier base de datos de este tipo de seguro.

7. BIBLIOGRAFÍA

Art. 1, Ley 50/1980, 8 de octubre, de Contrato de Seguro. (17 de Octubre de 1980). BOE, núm. 250. España: Gobierno de España. Ministerio de la presidencia, relaciones con las cortes e igualdad.

Art. 27, Ley 50/1980, 8 de octubre, de Contrato de Seguro. (17 de Octubre de 1980). BOE, núm. 250. España: Gobierno de España. Ministerio de la presidencia, relaciones con las cortes e igualdad.

Art. 83, Ley 50/1980, 8 de octubre, de Contrato de Seguro. (17 de Octubre de 1980). BOE, núm. 250. España: Gobierno de España. Ministerio de la presidencia, relaciones con las cortes e igualdad.

Avasant. (Mayo 2017). *La robótica aplicada al sector asegurador en España.*

Ayuntamiento de Madrid. (Mayo de 2019). *Portal Web del Ayuntamiento de Madrid*. Obtenido de https://www.madrid.es

Cameron, A. C and Trivedy, K.P. (1998). *Regression analysis of count data.* Cambridge University.

Comunidad de Madrid. (3 de Junio de 2019). Obtenido de Geoportal de la Comunidad de Madrid: https://idem.madrid.org/

DGSFP. (2017). *Informe 2017. Seguros y Fondos de Pensiones.*

Fundación MAPFRE. (Mayo de 2019). Obtenido de Seguros: https://segurosypensionesparatodos.fundacionmapfre.org

Hilbe, J.M. (2011). *Negative binomial regression.* Cambridge University.

ICEA. (Febrero de 2019). *Primas y otros datos. Evolución del Mercado Asegurador. Estadística a diciembre. Año 2018. Recuperado en Abril de 2019, de Servicio de estadísticas y estudios del sector asegurador en España ICEA https://www.icea.es*.

ICEA. (2019). Crecimiento interanual de primas y pólizas de negocio multirriesgo. Recuperado en Abril de 2019, de Servicio de estadísticas y estudios del sector asegurador en España ICEA https://www.icea.es.

Idealista. David Rey. (14 de Febrero de 2019). Ponencia Data Day 2019: De la Monetización a la Robotización en el Sector Inmobiliario. Madrid.

INE. (2015). *Datos obtenidos de la Encuesta de Presupuesto familiares*. Obtenido de https://www.ine.es

INE. (2019). *Encuesta Continua de Hogares*. Obtenido de Instituto Nacional de Estadística: http://www.ine.es

Ismael Gómez Schmidt. (2018). *satRday Santiago 2018. Una mirada al Soccer Analytics usando R.* Chile.

Judd C. M et al. (2009). Data analysis: a model comparison approach. Routledge.

María Elena Fernández Boza. (Julio de 2017). Métodos de Regresión para Count Data. Jaén.

Ministerio del Interior. (Mayo de 2019). *Gobierno de España. Ministerio del Interior*. Obtenido de Balances trimestralidades de criminalidad. : https://estadisticasdecriminalidad.ses.mir.es/

Período ABC. (2 de 6 de 2016). El invierno y agosto, los momentos preferidos por los ladrones de casas para robar. *ABC SOCIEDAD.*

RAE. (2019). Obtenido de Real Academia Española: https://dle.rae.es

Tor Bernhardsen. (2002). *Geographic Information Systems: An Introduction. 3rd Edition.* Norway.

UNESPA. (2014). *Los robos en los hogares madrileños.* Madrid.

Victor Olaya. (2014). *Sistema de información Geográfica.*

Wikipedia. (2019). Obtenido de Charles_Picquet: https://es.wikipedia.org

Wikipedia. (2019). Obtenido de John_Snow: https://es.wikipedia.org

Wikipedia. (2019). Obtenido de Seguro Multirriesgo: https://es.wikipedia.org

Wilkelman, R. (2008). *Enometric Analysis of Count Data, Fifth Edition.* Jaén: Springer.

Anexo A

Gráfico 2. Comparativa Ramos. Elaboración Propia

Tipo de ramo	2013	2014	2015	2016	2017
Ramo vida	46,06%	45,20%	45,19%	48,23%	45,94%
Ramo no vida	53,94%	54,80%	54,81%	51,77%	54,06%

Gráfico 3. Volumen Estimado de Primas por Ramo. Elaboración Propia

Ramo vida	Ramo no vida
28.994.769.567	35.382.070.066

Gráfico 4. Volumen Estimado Primas por Modalidad. Año 2018. Elaboración Propia

	2018
Automóviles	11.134.847.323,75
Multirriesgos	7.244.373.772,07
Salud	85.236.50.570,59
Otros No Vida	8479198399,2
Riesgo	4.721.001.071,78
Ahorro	24.273.768.495,10

	AÑO 2017	AÑO 2018	
Total Seguro Directo	**63.433.933.551**	**64.376.839.632**	**1,49%**
No Vida	**34.027.090.654**	**35.382.070.066**	**3,98%**
Automóviles	**10.923.292.612**	**11.134.847.324**	**1,94%**
Automóviles RC	5.716.640.205	5.849.933.952	2,33%
Automóviles Otras Garantías	5.206.652.407	5.284.913.372	1,50%
Multirriesgos	**6.964.662.950**	**7.244.373.772**	**4,02%**
Hogar	4.196.339.275	4.346.255.241	3,57%
Comercio	584.815.662	596.010.988	1,91%
Comunidades	872.102.393	898.482.125	3,02%
Industrias	1.233.634.002	1.327.164.585	7,58%
Otros	77.771.618	76.460.833	-1,69%
Salud	**8.068.725.719**	**8.523.650.571**	**5,64%**
Asistencia Sanitaria	7.100.835.254	7.532.218.654	6,08%
Reembolso	708.821.306	722.315.638	1,90%
Subsidios	259.069.159	269.116.278	3,88%
Total Resto No Vida	**8.070.409.373**	**8.479.198.399**	**5,07%**
Accidentes	1.114.331.043	1.151.947.134	3,38%
Asistencia	402.699.109	446.626.578	10,91%
Caución	63.178.565	78.889.209	24,87%
Crédito	570.332.327	584.360.880	2,46%
Decesos	2.276.991.267	2.367.626.085	3,98%
Defensa Jurídica	100.518.894	106.363.806	5,81%
Incendios	118.408.335	145.699.060	23,05%
Riesgos industriales	91.306.344	119.302.921	30,66%
Resto incendios	27.101.991	26.396.139	-2,60%
Otros daños a los bienes	1.105.704.097	1.223.165.694	10,62%
Avería maquinaria	107.253.928	126.495.962	17,94%
Equipos Electrónicos	53.205.322	54.993.854	3,36%
Montaje	10.387.691	14.446.909	39,08%
Robo	24.157.697	25.172.284	4,20%
Seguro decenal	28.190.092	32.257.111	14,43%
Todo riesgo construcción	47.661.347	58.608.227	22,97%
Resto Otros Daños a los bienes	834.848.021	911.191.347	9,14%
Pérdidas pecuniarias	350.905.980	376.688.655	7,35%
Responsabilidad civil	1.508.329.667	1.537.280.621	1,92%
Transportes	459.010.088	460.550.678	0,34%
Aviación	67.610.366	45.981.643	-31,99%
Marítimo	161.352.300	169.556.292	5,08%
Mercancías	230.047.421	245.012.743	6,51%
Vida	**29.406.842.897**	**28.994.769.567**	**-1,40%**
Riesgo	**4.205.550.075**	**4.721.001.072**	**12,26%**
Ahorro	**25.201.292.822**	**24.273.768.495**	**-3,68%**

Gráfico 5. Volumen Estimado de Primas Emitidas de Seguros Multirriesgo. Elaboración Propia

	2017	2018
Hogar	4.196.339.274,91	4.346.255.241,08
Comercio	584.815.661,71	596.010.987,8
Comunidades	872.102.393,49	898.482.125,41
Industrias	1.233.634.002,45	1.327.164.584,82
Otros	77.771.617,61	76.460.832,96

Anexo B

Tabla 23.Variables de Estudio

VARIABLE	TIPO DE VARIABLE	CODIFICACIÓN SI PROCEDE
Cod_Distrito	Dummy	01 - … - 21
Distritos	Dummy	Centro – Arganzuela – Retiro - Salamanca – Chamartín – Tetuán – Chamberí – Fuencarral – Moncloa – Latina – Carabanchel – Usera – Puente de Vallecas – Moratalaz – Ciudad Lineal – Hortaleza – Villaverde – Villa de Vallecas – Vicálvaro – San Blas - Barajas
Mes	Dummy	Enero - … - Diciembre
Año	CUANTITATIVA	
Robos_Viviendas	CUANTITATIVA	
Tasa_Robos	CUANTITATIVA	
Estacional_1	Dummy (Dicotómica)	$\begin{cases} 0 & \text{No Semana Santa} \\ 1 & \text{Semana Santa} \end{cases}$
Estacional_2	Dummy (Dicotómica)	$\begin{cases} 0 & \text{No Navidad} \\ 1 & \text{Navidad} \end{cases}$
Estacional_3	Dummy Dicotómica)	$\begin{cases} 0 & \text{No Vacaciones} \\ 1 & \text{Vacaciones} \end{cases}$
Superficie (Ha)	CUANTITATIVA	
Densidad_Pob. (Hab/Ha)	CUANTITATIVA	
Tasa_Absoluta_Paro	CUANTITATIVA	
Población_Española	CUANTITATIVA	
Población_Extranjera	CUANTITATIVA	
Renta_Media_Alquiler_€/m^2	CUANTITATIVA	
Precio_Vivienda_2ª_Mano_€/m^2	CUANTITATIVA	
Renta_Bruta_Disponible	CUANTITATIVA	
Nº Festivos Laborables	CUANTITATIVA	

Anexo C

Código Python

```
#IMPORTAR PAQUETES
#GRAFICOS
import plotly.graph_objs as go
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
init_notebook_mode(connected=True)
from plotly import tools

#TABLAS
import plotly.plotly as py
import plotly.graph_objs as go
import plotly.figure_factory as ff
```

```
##COMPARATIVA RAMOS
#DATOS RAMO VIDA
VIDA = go.Bar(
    x = ['2013', '2014', '2015', '2016', '2017'],
    y = [46.06, 45.20, 45.19, 48.23, 45.94],
    name = 'Ramo vida',
    text = ['46.06%', '45.20%', '45.19%', '48.23%', '45.94%'],
    marker = dict(color = 'rgb(175, 49, 35)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size = 12, color = 'white'))
#DATOS RAMO NO VIDA
NO_VIDA = go.Bar(
    x = ['2013', '2014', '2015', '2016', '2017'],
    y = [53.94, 54.80, 54.81, 51.77, 54.06],
    name = 'Ramo no vida',
    text = ['53.94%', '54.80%', '54.81%', '51.77%', '54.06%'],
    marker = dict(color = 'rgb(33, 75, 99)'), textposition = 'auto',
    textfont=dict(family = 'Time New Roman', size=12, color = 'white'))
## FUNCIÓN GRÁFICA
DATOS_GRAFICO_1 = [VIDA, NO_VIDA]

layout = go.Layout(
    barmode = 'group',
    yaxis = dict(title = '%',
                showgrid = True, showline = True,
                mirror = 'ticks', gridcolor = '#bdbdbd',
                gridwidth = 1, linecolor = '#636363', linewidth = 1),
    xaxis = dict(title = 'Periodos', tickangle = 0,
                showgrid = True, showline = True,
                mirror = 'ticks', gridcolor = '#bdbdbd',
                gridwidth = 1, linecolor = '#636363', linewidth = 1),
#LEYENDA
    legend = dict(orientation = "h"),
    font = dict(family = 'Times New Roman', size = 14))
GRAFICO_1 = go.Figure(data = DATOS_GRAFICO_1, layout = layout)
```

```
##VOLUMEN ESTIMADO DE PRIMAS POR RAMO
RAMOS_SECTORES = {
    "data": [{"values": [28994769566.8894, 35382070065.5877],
              "labels": ['Ramo Vida', 'Ramo No Vida'],
              "marker": {'colors': ['rgb(175, 49, 35)', 'rgb(33, 75, 99)']},
              "textposition": "inside",
              "textfont": dict(family = 'Time New Roman', size = 14, color = 'white'),
              "domain": {"column": 0},
              # "textinfo":'percent',
              "name": "Tipo de Ramos",
              "hoverinfo": "label+percent+name",
              "type": "pie"}],
    "layout": {"grid": {"rows": 1, "columns": 1}}}
```

```
##VOLUMEN ESTIMADO DE PRIMAS POR MODALIDAD
PRIMAS_MODALIDAD = {
  "data": [
    {"values": [11134847323.75, 7244373772.07, 8523650570.59,
    8479198399.2, 4721001071.78, 24273768495.10],
      "labels": ["Automóviles", "Multirriesgos", "Salud",
      "Otros No vida", "Riesgo", "Ahorro"],
     "marker": {'colors': ['rgb(33, 75, 99)', 'rgb(79, 129, 102)',
                'rgb(151, 179, 100)', 'rgb(175, 49, 35)', 'rgb(36, 73, 147)','rgb(146, 123,
21)']},
      "textposition":"inside",
     "textfont":dict(family='Time New Roman',
                     size=14,
                     color='white'),
      "domain": {"column": 1},
      "hoverinfo":"label+percent+name",
      "hole": .4, "type": "pie"}],
  "layout": {"grid": {"rows": 1, "columns": 1}}}
```

```
##VOLUMEN ESTIMADO DE PRIMAS SEGUROS MULTIRRIESGO
MULTIRRIESGO = {
  "data": [{"values": [4196339274.91, 584815661.71, 872102393.49,
                       1233634002.45, 77771617.61],
      "labels": ["Hogar 2017", "Comercio 2017", "Comunidades 2017",
      "Industrias 2017", "Otros 2017"],
      "marker": {'colors': ['rgb(33, 75, 99)', 'rgb(79, 129, 102)',
                'rgb(151, 179, 100)', 'rgb(175, 49, 35)', 'rgb(36, 73, 147)']},
      "textposition":"inside",
      "textfont":dict(family='Time New Roman',
                      size=14,
                      color='white'),
      "domain": {"column": 0},

      "name": "Tipo de Ramos",
      "hoverinfo":"label+percent+name",
      "hole": .4, ###TAMAÑO CIRCULO
      "type": "pie"},
```

```
{"values": [4346255241.08, 596010987.8, 898482125.41,
    1327164584.82, 76460832.96],
      "labels": ["Hogar 2018", "Comercio 2018", "Comunidades 2018",
      "Industrias 2018", "Otros 2018"],
      "marker": {'colors': ['rgb(33, 75, 99)', 'rgb(79, 129, 102)',
                'rgb(151, 179, 100)', 'rgb(175, 49, 35)', 'rgb(36, 73, 147)']},
      "textposition":"inside",
      "textfont":dict(family='Time New Roman', size=14,
                      color='white'),
      "domain": {"column": 1},
      "hoverinfo":"label+percent+name",
      "hole": .4, "type": "pie"}],
      "layout": {"grid": {"rows": 1, "columns": 2},
       "annotations": [{"font": {"size": 22},
                "showarrow": False,
                "text": "2017",
                "x": 0.18, "y": 0.5},
                {"font": {"size": 22},
                "showarrow": False,
                "text": "2018",
                "x": 0.825, "y": 0.5}]}}
```

```
##CRECIMIENTO TOTAL INTERANUAL NEGOCIO MULTIRRIESGO
#DATAFRAMES
TOTAL_CREC_MULT = {'Fecha': ['Mar 2017', 'Jun 2017', 'Sept 2017', 'Dic 2017', 'Mar 2018',
'Jun 2018', 'Sept 2018', 'Dic 2018'],
    'Primas': [1.85, 2.53, 2.35, 2.30, 2.72, 3.68, 3.60, 3.95],
    'Polizas': [2.30, 1.67, 1.79, 1.85, 1.34, 2.01, 1.91, 1.93]}
TOTAL_CRECIMIENTO = pd.DataFrame(TOTAL_CREC_MULT)
TOTAL_CREC_MULT_2017 = {'Fecha': ['Mar', 'Jun', 'Sept', 'Dic'],
                        'Primas': [1.85, 2.53, 2.35, 2.30],
                        'Polizas': [2.30, 1.67, 1.79, 1.85]}
TOTAL_2017 = pd.DataFrame(TOTAL_CREC_MULT_2017)
TOTAL_CREC_MULT_2018 = {'Fecha': ['Mar', 'Jun', 'Sept', 'Dic'],
                        'Primas': [2.72, 3.68, 3.60, 3.95],
                        'Polizas': [1.34, 2.01, 1.91, 1.93]}
TOTAL_2018 = pd.DataFrame(TOTAL_CREC_MULT_2018)
```

```
##POLIZAS Y PRIMAS
PRIMAS_2017_TOTAL = go.Bar(
    x = TOTAL_2017.Fecha,
    y = TOTAL_2017.Primas,
    name = 'Primas 2017', text = ['1,85%', '2,53%', '2,35%', '2,30%'],
    marker = dict(color = 'rgb(151, 179, 100)'), textposition='auto',
    textfont = dict(family = 'Time New Roman', size = 12, color = 'white'))

POLIZAS_2017_TOTAL = go.Bar(
    x = TOTAL_2017.Fecha,
    y = TOTAL_2017.Polizas,
    name = 'Polizas 2017', text = ['2,30%', '1,67%', '1,79%', '1,85%'],
    marker = dict(color = 'rgb(33, 75, 99)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size = 12,  color = 'white'))
```

```
PRIMAS_2018_TOTAL = go.Bar(
    x = TOTAL_2018.Fecha,
    y = TOTAL_2018.Primas,
    name = 'Primas 2018',
    text = ['2,72%', '3,68%', '3,60%', '3,95%'],
    marker = dict(color = 'rgb(79, 129, 102)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size=12, color = 'white'))

POLIZAS_2018_TOTAL = go.Bar(
    x = TOTAL_2018.Fecha,
    y = TOTAL_2018.Polizas,
    name ='Polizas 2018',
    text = ['1,34%', '2,01%', '1,91%', '1,93%'],
    marker = dict(color = 'rgb(175, 49, 35)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size = 12, color = 'white'))

POLIZAS_PRIMAS = [PRIMAS_2017_TOTAL, POLIZAS_2017_TOTAL, PRIMAS_2018_TOTAL,
POLIZAS_2018_TOTAL]

layout = go.Layout(barmode='stack',
                   yaxis = dict(title = '%', showgrid = True, showline = True,
                                mirror = 'ticks', gridcolor = '#bdbdbd',
                                gridwidth = 1, linecolor = '#636363', linewidth = 1),
                   xaxis = dict(tickangle = 0, showgrid = True, showline = True,
                                mirror = 'ticks', gridcolor = '#bdbdbd',
                                gridwidth=1, linecolor='#636363', linewidth = 1),
                                font=dict(family = 'Times New Roman', size = 12),
                   legend = dict(orientation = "h"))
GRAF_CRECIMIENTO = go.Figure(data=POLIZAS_PRIMAS, layout=layout)
```

```
##PRIMAS
PRIMAS_2017 = go.Bar(
    x = TOTAL_2017.Fecha, y = TOTAL_2017.Primas,
    name = 'Primas 2017', text = ['1,85%', '2,53%', '2,35%', '2,30%'],
    marker = dict(color='rgb(151, 179, 100)'), textposition = 'auto',
    textfont = dict(family='Time New Roman', size = 12, color = 'white'))

PRIMAS_2018 = go.Bar(
    x = TOTAL_2018.Fecha, y = TOTAL_2018.Primas,
    name = 'Primas 2018', text = ['2,72%', '3,68%', '3,60%', '3,95%'],
    marker = dict(color = 'rgb(79, 129, 102)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size=12, color = 'white'))
PRIMAS = [PRIMAS_2017, PRIMAS_2018]
layout = go.Layout(barmode='group',
                   yaxis = dict(title='%', showgrid = True, showline = True,
                                mirror='ticks', gridcolor='#bdbdbd',
                                gridwidth = 1, linecolor='#636363', linewidth = 1),
                   xaxis = dict(tickangle=0,showgrid = True, showline = True,
                                mirror = 'ticks', gridcolor ='#bdbdbd',
                                gridwidth = 1, linecolor = '#636363',
                                font=dict(family='Times New Roman', size = 12), linewidth=1),
                   legend = dict(orientation = "h"))
GRAF_PRIMAS = go.Figure(data = PRIMAS, layout = layout)
```

```
##POLIZAS
POLIZAS_2017 = go.Bar(
    x = TOTAL_2017.Fecha,
    y = TOTAL_2017.Polizas,
    name = 'Pólizas 2017',
    text = ['2,30%', '1,67%', '1,79%', '1,85%'],
    marker = dict(color = 'rgb(33, 75, 99)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size = 12, color = 'white'))
POLIZAS_2018 = go.Bar(
    x = TOTAL_2018.Fecha,
    y = TOTAL_2018.Polizas,
    name = 'Polizas 2018',
    text = ['1,34%', '2,01%', '1,91%', '1,93%'],
    marker = dict(color='rgb(175, 49, 35)'), textposition = 'auto',
    textfont = dict(family = 'Time New Roman', size = 12, color = 'white'))
POLIZAS = [POLIZAS_2017, POLIZAS_2018]

layout = go.Layout(barmode='group',
                   yaxis=dict(title = '%', showgrid = True, showline = True,
                              mirror = 'ticks', gridcolor = '#bdbdbd',
                              gridwidth = 1, linecolor = '#636363', linewidth = 1),
                   xaxis = dict(tickangle = 0, showgrid = True, showline = True,
                              mirror = 'ticks', gridcolor = '#bdbdbd',
                              gridwidth = 1, linecolor = '#636363', linewidth = 1),
                   font = dict(family = 'Times New Roman', size = 12),
                   legend = dict(orientation = "h"))
GRAF_POLIZAS = go.Figure(data = POLIZAS, layout=layout)
#UNIÓN GRÁFICOS
UNION_GRAFICOS = tools.make_subplots(rows = 3, cols = 1)
fig.append_trace(POLIZAS_2017, 1, 1)
fig.append_trace(POLIZAS_2018, 1, 1)
fig.append_trace(PRIMAS_2017, 2, 1)
fig.append_trace(PRIMAS_2018, 2, 1)
fig.append_trace(PRIMAS_2017_TOTAL, 3, 1)
fig.append_trace(POLIZAS_2017_TOTAL, 3, 1)
fig.append_trace(PRIMAS_2018_TOTAL, 3, 1)
fig.append_trace(POLIZAS_2018_TOTAL, 3, 1)
```

```
##CRECIMIENTO INTERANUAL PRIMAS MULTIRRIESGO
labels = ['Hogar', 'Comercio', 'Comunidades', 'Industrial', 'Otros']
colors = ['rgb(146, 123, 21)', 'rgb(115,115,115)', 'rgb(49,130,189)', 'rgb(175, 51, 21)',
'rgb(69, 139, 0)']

mode_size = [8, 8, 8, 8, 8]
line_size = [2, 2, 2, 2, 2]
FECHAS = [['Mar 2017', 'Jun 2017', 'Sept 2017', 'Dic 2017', 'Mar 2018', 'Jun 2018', 'Sept
2018', 'Dic 2018'], ['Mar 2017', 'Jun 2017', 'Sept 2017', 'Dic 2017', 'Mar 2018', 'Jun
2018', 'Sept 2018', 'Dic 2018'], ['Mar 2017', 'Jun 2017', 'Sept 2017', 'Dic 2017', 'Mar
2018', 'Jun 2018', 'Sept 2018', 'Dic 2018'], ['Mar 2017', 'Jun 2017', 'Sept 2017', 'Dic
2017', 'Mar 2018', 'Jun 2018', 'Sept 2018', 'Dic 2018'], ['Mar 2017', 'Jun 2017', 'Sept
2017', 'Dic 2017', 'Mar 2018', 'Jun 2018', 'Sept 2018', 'Dic 2018'],]
```

```
VARIABLES = [[3.09, 2.998, 2.932, 2.936, 2.735, 3.341, 3.271,3.49],
    [2.23, 1.797, 1.561, 1.781, 2.739, 2.396, 2.178, 1.92],
    [2.52, 2.089, 2.035, 2.270, 0.806, 2.317, 2.673, 3.03],
    [-2.40, 1.833, 0.969, 0.578, 4.190, 6.609, 6.560, 7.59],
    [-1.48, -1.426, -0.599, -4.107, -1.032, -1.926, -4.908, -2.55],]
MULTI = []
for i in range(0, 5):
MULTI.append(go.Scatter(
        x = FECHAS[i],
        y = VARIABLES[i],
        mode = 'lines+markers',
        line = dict(color=colors[i], width = line_size[i]), connectgaps = True, ))
MULTI.append(go.Scatter(
    x = [FECHAS[i][0], FECHAS[i][7]],
    y = [VARIABLES[i][0], VARIABLES[i][7]],
    mode = 'markers', marker = dict(color = colors[i], size = mode_size[i])))

layout = go.Layout(
xaxis = dict(showline = True,
        showgrid = False,
        showticklabels = True,
        linecolor = 'rgb(204, 204, 204)', linewidth = 2,
        ticks = 'outside', tickcolor = 'rgb(204, 204, 204)',
        tickwidth = 2, ticklen = 5,tickfont = dict(family = 'Times New Roman',
                    size = 12, color = 'rgb(82, 82, 82)')),
yaxis = dict(showgrid = False,
        zeroline = True,
        showline = False,
        showticklabels = False),
        autosize = False,margin = dict(autoexpand = False,l = 100, r = 20, t = 4, b = 25),
        showlegend = False)
annotations = []
# ETIQUETAS
for y_ MULTI, label, color in zip(y_data, labels, colors):
    annotations.append(dict(xref= 'paper', x = 0.05, y = y_ MULTI [0],
                            xanchor = 'right', yanchor = 'middle',
                            text=label + ' {}%'.format(y_ MULTI [0]),
                            font=dict(family = 'Times New Roman',
                                      size = 10), showarrow = False))

annotations.append(dict(xref = 'paper', x=0.95, y = y_ MULTI [7],
                            xanchor = 'left', yanchor = 'middle',
                            text='{}%'.format(y_trace[7]),
                            font=dict(family = 'Times New Roman',
                                      size = 10), showarrow = False))
layout['annotations'] = annotations
GRAF_INTER = go.Figure(data = MULTI, layout = layout)
```

```
#TABLAS
##EL SEGURO EN LA ECONOMIA ESPAÑOLA
TABLA_1 = go.Table(
  header = dict(
    values = [[''], ['<b>2013'],['<b>2014'],
              ['<b>2015'],['<b>2016'],['<b>2017']],
    line = dict(color = '#506784'),
    fill = dict(color = 'rgb(33, 75, 99)'),
    align = ['left','center'],
    font = dict(family = 'Times New Roman', color = 'white', size = 14)),
    cells = dict(
    values = [[['PDB'],['PIB a p.m.'],
               ['PDB/PIB'], ['PDB/habitantes']],
              [['56.263'], ['1.025.634'], ['5,49%'], [1.194]],
              [['56.016'], ['1.037.025'], ['5,40%'], [1.198]],
              [['57.073'], ['1.075.639'], ['5,31%'], [1.224]],
              [['64.920'], ['1.118.522'], ['5,80%'], [1.394]],
              [['64.514'], ['1.163.662'], ['5,54%'], [1.385]]],
    line = dict(color = '#506784'),
    align = ['left','center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
DATOS_TABLA_1 = [TABLA_1]
TABLA_ECONOMIA = dict(data = DATOS_TABLA_2, layout = layout)
```

```
##PRIMAS DEVENGADAS BRUTAS Y VARIACIÓN TOTAL
TABLA_2 = go.Table(
  header = dict(values = [[''], ['<b>2013'],['<b>2014'],
              ['<b>2015'],['<b>2016'],['<b>2017']],
    line = dict(color = '#506784'),
    fill = dict(color = 'rgb(33, 75, 99)'),
    align = ['left','center'],
    font = dict(family = 'Times New Roman', color = 'white', size = 14)),
    cells = dict(values = [[['PDB'],['Ramo vida'],
               ['Ramo no vida'], ['PDB a p.m.'], ['Variación total'],
               ['Variación vida'], ['Var. no vida'], ['Var. PIB a p.m.']],
    [['56.263'], ['25.913'], ['30.350'], ['1.025.634'],['-2,57%'], ['-2,98%'], ['-2,22%'],
['-1,36%']],
    [['56.016'], ['25.321'], ['30.695'], ['1.037.025'], ['-0,44%'], ['-2,29%'], ['1,14%'],
['1,11%']],
    [['57.073'], ['25.791'], ['31.282'], ['1.075.639'], ['1,89%'], ['1,86%'], ['1,91%'],
['3,72%']],
    [['64.920'], ['31.309'], ['33.612'], ['1.118.522'], ['13,75%'], ['21,39%'], ['7,45%'],
['3,99%']],
    [['64.514'], ['29.639'], ['34.875'], ['1.163.662'], ['-0,63%'], ['-5,33%'], ['3,76%'],
['4,04%']]], line = dict(color = '#506784'), align = ['left','center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
DATOS_TABLA_2 = [TABLA_2]
TABLA_PRIMAS = dict(data = DATOS_TABLA_2, layout = layout)
```

```
#TABLAS ANEXO
##TABLA 1 ANEXO: COMPARATIVA RAMOS
TABLA_ANEXO_A_FIG_1 = go.Table(
  header = dict(values = [['<b>Tipo de ramo'], ['<b>2013'],['<b>2014'],
              ['<b>2015'], ['<b>2016'], ['<b>2017']],
    line = dict(color = '#506784'),
    fill = dict(color = 'rgb(33, 75, 99)'),
    align = ['left','center'],
    font = dict(family = 'Times New Roman', color = 'white', size = 14)),
    cells = dict(values = [[['Ramo vida'],['Ramo no vida']],
              [['46,06%'], ['53,94%']],
              [['45,20%'], ['54,80%']],
              [['45,19%'], ['54,81%']],
              [['48,23%'], ['51,77%']],
              [['45,94%'], ['54,06%']]],
    line = dict(color = '#506784'),
    align = ['left','center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
DATOS_TABLA_ANEXO_A_FIG_1 = [TABLA_ANEXO_A_FIG_1]
TABLA_ANEXO_A_FIG_1 = dict(data = DATOS_TABLA_ANEXO_A_FIG_1, layout = layout)
```

```
##TABLA 2 ANEXO A: VOLUMEN ESTIMADO DE PRIMAS POR RAMO
TABLA_ANEXO_A_FIG_2 = go.Table(
  header = dict(
    values = [['<b>Ramo vida'], ['<b>Ramo no vida']],line = dict(color = '#506784'),
    fill =  dict(color = 'lightgrey'), align = ['center'],
    font = dict(family = 'Times New Roman', color = 'black',size =  14)),
    cells = dict(
    values = [['28.994.769.567'], ['35.382.070.066']],
    line = dict(color = '#506784'),
    align = ['center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
layout = dict(width=500, height=500)
DATOS_TABLA_ANEXO_A_FIG_2 = [TABLA_ANEXO_A_FIG_2]
TABLA_ANEXO_A_FIG_2 = dict(data = DATOS_TABLA_ANEXO_A_FIG_2, layout = layout)
```

```
##TABLA 3 ANEXO A: VOLUMEN ESTIMADO DE PRIMAS POR MODALIDAD
TABLA_ANEXO_A_FIG_3 = go.Table(
  header = dict(values = [[''], ['<b>2018']],
    line = dict(color = '#506784'), fill = dict(color = 'grey'), align = ['center'],
    font = dict(family = 'Times New Roman', color = 'white', size = 14)),
    cells = dict(values = [[['Automóviles'], ['Multirriesgos'], ['Salud'],
          ['Otros No   Vida'], ['Riesgo'], ['Ahorro']],
          [['11.134.847.323,75'], ['7.244.373.772,07'], ['85.236.50.570,59'],
         ['8479198399,2'], ['4.721.001.071,78'], ['24.273.768.495,10']]],
    line = dict(color = '#506784'), align = ['center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
layout = dict(width=500, height=500)
DATOS_TABLA_ANEXO_A_FIG_3 = [TABLA_ANEXO_A_FIG_3]
TABLA_ANEXO_A_FIG_3 = dict(data = DATOS_TABLA_ANEXO_A_FIG_3, layout = layout)
```

```
## TABLA 4 ANEXO A: VOLUMEN ESTIMADO DE PRIMAS SEGURO MULTIRRIESGO
TABLA_ANEXO_A_FIG_4 = go.Table(
  header = dict(
    values = [[''], ['<b>2017'], ['<b>2018']],
    line = dict(color = '#506784'),
    fill = dict(color = 'grey'),
    align = ['center'],
    font = dict(family = 'Times New Roman', color = 'white', size = 14)),
    cells = dict(
    values = [[['Hogar'], ['Comercio'], ['Comunidades'],
               ['Industrias'], ['Otros']],
              [['4.196.339.274,91'], ['584.815.661,71'],
        ['872.102.393,49'], ['1.233.634.002,45'],
                ['77.771.617,61']], [['4.346.255.241,08'],
                ['596.010.987,8'], ['898.482.125,41'],
                ['1.327.164.584,82'], ['76.460.832,96']]],
    line = dict(color = '#506784'),
    align = ['center'], height = 25,
    font = dict(family = 'Times New Roman', color = 'black', size = 14)))
layout = dict(width=500, height=500)
DATOS_TABLA_ANEXO_A_FIG_4 = [TABLA_ANEXO_A_FIG_4]
TABLA_ANEXO_A_FIG_4 = dict(data = DATOS_TABLA_ANEXO_A_FIG_4, layout = layout)
```

```
## PAQUETES BÁSICOS
import pandas as pd
import numpy as np
```

```
## PAQUETES CORRELACIONES
import matplotlib
import matplotlib.pyplot as plt
import os
import numpy as np
import pandas as pd
import seaborn as sns
import StringIO
```

```
## PAQUETES GRAFICAS DISTRIBUCIONES
import scipy.stats as st
import matplotlib.pyplot as plt
import scipy.interpolate as interpolate
```

```
#MATRIZ CORRELACIONES
mask = np.zeros_like(var_numericas.corr(), dtype = np.bool)
mask[np.triu_indices_from(mask)] = True
f, ax = plt.subplots(figsize=(12, 9))
corr = sns.heatmap(var_numericas.corr(),
                   mask = mask, cmap = "YlGnBu",
                   linewidths = 1,
                   cbar_kws={"shrink": .5},
                   annot = True,
                   annot_kws = {"size": 10})
```

```
##DISTRIBUCIÓN POISSON
plt.figure(figsize = (10,6))
mu =  4
poisson = st.poisson(mu)
x = np.arange(poisson.ppf(0.01),poisson.ppf(0.99))
fmp = poisson.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))
mu =  5
poisson = st.poisson(mu)
x = np.arange(poisson.ppf(0.01),poisson.ppf(0.99))
fmp = poisson.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))

mu =  6
poisson = st.poisson(mu)
x = np.arange(poisson.ppf(0.01),poisson.ppf(0.99))
fmp = poisson.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))

plt.ylabel("Probabilidad")
plt.xlabel("Conteos")
plt.legend(['mu = 5','mu = 4','mu = 6'])
plt.show()
```

```
##DISTRIBUCIÓN BINOMIAL NEGATIVA P CONSTANTE (PARAMETRO DE FORMA)

plt.figure(figsize=(10,6))
n =  15
p = 0.6
nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))
n =  6
p = 0.6

nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s=0, k=4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))

n = 10
p = 0.6
nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)

t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))
plt.ylabel("Probabilidad")
plt.xlabel("Conteos")
plt.legend(['n = 15 p = 0,6','n = 6 p = 0,6','n = 10 p = 0,6'])
plt.show()
```

```
##DISTRIBUCIÓN BINOMIAL NEGATIVA N CONSTANTE

plt.figure(figsize = (10,6))
n =  6
p = 0.75
nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))

n =  6
p = 0.7
nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s=0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))

n =  6
p = 0.8
nbinom = st.nbinom(n,p)
x=np.arange(nbinom.ppf(0.01),nbinom.ppf(0.99))
fmp = nbinom.pmf(x)
t, c, k = interpolate.splrep(x, fmp, s = 0, k = 4)
spline = interpolate.BSpline(t, c, k, extrapolate = False)
N = 100
xmin, xmax = x.min(), x.max()
xx = np.linspace(xmin, xmax, N)
plt.plot(xx, spline(xx))
plt.ylabel("Probabilidad")
plt.xlabel("Conteos")
plt.legend(['n = 6 p = 0,8','n = 5 p = 0,7','n = 5 p = 0,75'])
plt.show()
```

```
##PAQUETES GLM
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.graphics.gofplots import ProbPlot
from sklearn.preprocessing import LabelEncoder, StandardScaler, OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

```
##TRANSFORMACION LOGARITMICA
x['Renta_Media_Alquiler_Em2'] = np.log(x['Renta_Media_Alquiler_Em2'])
x['Precio_Vivienda_Segunda_Mano_Em2'] = np.log(x['Precio_Vivienda_Segunda_Mano_Em2'])
x['Renta_Bruta_Disponible'] = np.log(x['Renta_Bruta_Disponible'])
```

```
##ESTIMACIÓN MODELO I

Modelo_1 = smf.glm(formula='Tasa_Robos ~ C(DISTRITOS) + C(MES) + Dias_Festivos_Laborales \
                    + C(Estacional_2) + C(Estacional_3)+ C(Estacional_1) +  Paro \
                     + Renta_Media_Alquiler_Em2 + Precio_Vivienda_Segunda_Mano_Em2'
                   ,data = df , offset = log_exposure, family =
sm.families.Poisson(link=sm.families.links.log))
Mod_1 = Modelo_1.fit()
print(Mod_1.summary2())
```

```
##MEDIDAS BONDAD DE AJUSTE
print('AIC:', Mod_1.aic)
print('BIC:', Mod_1.bic)

POISSON_2 = sum((Y_test-y_pred)**2)
RMSE_MOD_2 = math.sqrt(POISSON_2)
RMSE_MOD_2
```

```
##EVALUAR MODELO
y = dummy['Tasa_Robos']
x_1= dummy.drop(['COD_DISTRITO','Cobertura_robos_Domicilios','Pobl_Esp_Total',
             'Renta_Bruta_Disponible','Periodo','Pob_Extranjera_Total','Superficie_Ha_Mes',
                 'Tasa_Robos','Poblacion','Densidad_Poblacion'], axis = 1)
X_train, X_test, Y_train, Y_test = train_test_split(x_1, y, test_size = 0.2, random_state =
0)
regressor = LinearRegression()
regressor.fit(X_train, Y_train)
y_pred = regressor.predict(X_test)
print ('Podemos comparar nuestra predicción con nuestra Y_test para ver si nuestro modelo
tiene una buena performance: ')
performance = pd.DataFrame({"Y_test": Y_test, "Y_pred": y_pred, "Diferencia %":(abs(Y_test-
y_pred)/Y_test)*100 })
Valores = performance.sort_values(by=['Diferencia %'])
display(Valores[:10])
lasso = Lasso()
visualizer = PredictionError(lasso, alpha = 0.95, point_color = 'dodgerblue',
                             bestfit = True, line_color = 'mediumblue')
visualizer.fit(X_train, Y_train)
visualizer.score(X_test, Y_test)
g = visualizer.poof()
```