Anales
del Instituto de Actuarios Españoles

4ª Época, Número 30, Año 2024/147-168
DL M-3160-1961 – ISSN 0534-3232 – eISSN 2531-2308

HOW TO USE PUBLIC-PRIVATE DATABASES IN INSURANCE RISK MANAGEMENT: GEOGRAPHY, CLIMATE AND PEOPLE IN MOTOR INSURANCE

CÓMO USAR BASES DE DATOS PÚBLICO-PRIVADAS EN LA GESTIÓN DE RIESGOS ASEGURADORES: GEOGRAFÍA, CLIMA Y PERSONAS EN EL SEGURO DE AUTOMÓVILES

Luis Enrique Cespedes Coimbra
Universitat de Barcelona Business School. Barcelona, España.
ORCID: https://orcid.org/0009-0002-6491-7830
Lcespeco7@alumnes.ub.edu
(Autor para correspondencia)

Mercedes Ayuso Gutiérrez
Departamento de Econometría, Estadística y Economía Aplicada. Universitat de Barcelona. Barcelona, España.
ORCID: https://orcid.org/0000-0001-6127-4572
mayuso@ub.edu

Miguel Ángel Santolino Prieto
Departamento de Econometría, Estadística y Economía Aplicada. Universitat de Barcelona. Barcelona, España.
ORCID: https://orcid.org/0000-0002-0286-3673
msantolino@ub.edu

Reception date: July 13th 2024
Acceptance date: December 2nd 2024

ABSTRACT

This work focuses on the use of public information sources in the application of relational models in insurance companies, for a better understanding of risks and assisting decision-making in new sustainability environments. Firstly, we propose using Eurostat's degree of urbanization methodology to group motor claims or policies into potentially more homogeneous categories in the insurance sector (urban / suburban / rural) for segmentation and analysis. Secondly, we analyze how insurance companies can use local weather information in conjunction with the degree of urbanization to model the number of motor claims in a specific geographic area. Finally, we apply relational models to databases with anonymized information on passengers in traffic accidents provided by the Spanish General Traffic Directorate for the purpose of better defining the characteristics of the claim based on the profile of the people inside the vehicle. It is about knowing, for example, the profile of the passengers in vehicles driven by elderly people, also in conjunction with sex and the geographical area. Insurance companies know the enormous potential of data analytics and must focus on the search for relationships using information that may be dispersed in multiple databases, including those that are for public use and that can facilitate the homogenization and comparison of results, together to the design of preventive and risk management policies. We also include the R codes making them available to the insurance sector and academia for use.

Keywords: Data analytics, Relational models, Sustainability

DOI 10.26360/2024_08

RESUMEN

Este trabajo se centra en la utilización de fuentes de información pública en la aplicación de modelos relacionales en las entidades aseguradoras, para el mejor conocimiento de las características de los riesgos y asistir a la toma de decisiones en nuevos entornos de sostenibilidad. Primero, proponemos utilizar la metodología de grado de urbanización de Eurostat para agrupar siniestros o pólizas de automóviles en categorías potencialmente más homogéneas en el sector asegurador (urbano / suburbano / rural) para su segmentación y análisis. Segundo, analizamos como las compañías aseguradoras pueden utilizar información climatológica local conjuntamente con el grado de urbanización para modelizar el número de siniestros de automóviles en una zona geográfica específica. Finalmente, aplicamos modelos relacionales a bases de datos con información anonimizada de pasajeros en accidentes de tráfico proporcionadas por la Dirección General de Tráfico de España con el objetivo de definir mejor las características de los siniestros en función del perfil de las personas que se encuentran dentro del vehículo. Se trata de conocer, por ejemplo, el perfil de los pasajeros de vehículos conducidos por personas mayores, también en relación con el sexo y la zona geográfica. Las compañías aseguradoras conocen la enorme potencialidad del análisis de datos y deben apostar por la búsqueda de relaciones usando información que puede estar dispersa en múltiples bases de datos, incluyendo aquella que es de uso público y que puede facilitar la homogeneización y comparación de resultados, junto al diseño de políticas preventivas y de gestión de riesgos. Incluimos los códigos en R poniéndolos a disposición del sector asegurador y de la academia para su uso.

Palabras clave: Análisis de datos, Modelos relacionales, Sostenibilidad

1. INTRODUCTION

Data analysis has been at the core of the insurance industry since its inception. Insurance companies are ongoing an arms race to understand and apply advances in data science (Denuit et al., 2020; Wüthrich & Merz, 2023). Data science is a field of applied mathematics and statistics that extract information based on large amounts of complex data or big data. Data volume has exponentially grown in last years in the world. As more and more diverse data sources become available to insurance companies, techniques to link different databases to extract useful information for the insurance business become more important. However, there are cost-effective methods that enhance the understanding and pricing of risks that are not fully taken advantage of yet. One of them is the use of relational databases with internal private data and public available one.

In general, a data model is the formal way of expressing data relationships to a database management system (DBMS). The relational data model was introduced in 1970 by Edgar Frank Codd (1970). This model describes the world as a collection of inter-related tables, named relations (Watt & Eng, 2014). Databases that adhere to a relational data model are named relational databases. Therefore, a relational database is a database whose logical structure is made up of a collection of relations (Harrington, 2016). Relational databases work with base tables, i.e., actually stored tables, and virtual tables, which are a product of relational operations and only exist in main memory. Their structure is registered in a data dictionary or catalogue, which mirror data storage relations. The data within the data dictionary are referred to as metadata.

The aim of this study is to show that insurers have access to alternative data sources that are useful in pricing and risk management. We claim that relational models that integrate those data sources with internal data can be used by insurers in their risk analysis. There is abundant literature that provides interesting insights by combining various data sources with a relational model. Different aspects of mobility have been analyzed, all of which are of interest to insurance companies. There are several reviews that include research studies that have used a relational model to integrate the data. Ziakopoulos and Yannis (2020) wrote a literary review of spatial analysis approaches on road safety, where one can find studies that combine different data sources such as Liu & Sharma (2018), Moeinaddini et al. (2014) and Alarifi et al. (2017), as well as Ziakopoulos (2024) recently published research. In Zheng et al. (2021) we can find a review of studies modelling traffic conflicts that combine various data

sources, including among others Xie et al. (2019) and Zheng et al. (2019). Finally, Wang et al. (2013) examined the impact of traffic and road characteristics, and referenced some research studies that combined alternative data sources such as Haynes et al. (2008) and Lord et al. (2005).

To illustrate the usefulness of leveraging public data in combination with private data in a relational data model, this paper will show three fields where they can be used for a better understanding of risks. The first application focuses on attributing an urbanization degree to the claims and/or the policyholders, specifically the Eurostat methodology, in line with research that have leveraged ZIP codes to study the relationship between crash characteristics and those injured (Clark & Cushing, 2004; Lee et al., 2014; Lerner et al., 2001). The second application uses climatic information from AEMET (Spanish Meteorology Agency) to model the number of claims in a municipality (AEMET, 2024). Finally, we examine the characteristics of all occupants inside crashed vehicles, with the focus on the severity of the injuries of the passengers involved in a motor crash. Note that occupants are defined as all persons who were in the vehicle at the time of the accident, and passengers as the persons other than the drivers who were in the vehicle. Our idea is to obtain a more accurate estimate of the total bodily injury (BI) cost associated with an accident based on the profile of the driver and the expected passengers' profiles accompanying him.

This paper is structured into five sections. First, the methodology of relational databases is presented. Then, the three proposed applications are discussed. In the first one, we use relational data models in spatial analysis, specifically in segmenting geographical locations based on the European standard urbanization degree categorization (also considering the relevance of the analysis of geographical areas and their population structure). In the second application, we show how to use files with georeferencing raster climatic data to model the number of claims in a geographical area. Finally, relational databases are presented as a tool to link more data sets with the aim of better understanding the characteristics of occupant BI (driver plus passengers). The paper ends with the main conclusions on the relevance of using relational models in the insurance field. We also include the R codes making them available to the insurance sector and academia for use.

2. RELATIONAL DATABASES

Methodologically speaking, it is necessary to distinguish different components in a relational model, see: a relation, also named table, defined as a subset of the Cartesian product of a list of domains characterized by a name (Wijnen et al., 2019); its columns or attributes; its domains, i.e., the set of permissible values a column can contain; and its rows or tuples, where each one represents a group of related data values. These relation's rows and columns have some special properties:

- For columns: each one must have a unique name, its values must be drawn for only one domain, and viewing it in any order must not affect the meaning of the data.
- For rows: there must not be duplicate rows, there can only be one value at the intersection of a column and a row, and viewing rows in any order must not be affected the meaning of the data.

Each table row is identified with a primary key, a value that uniquely identifies a specific row, stored in a column. If well specified, with unique primary keys and no null keys, only the table and column names, and the primary key of the row suffice to retrieve any specific data.

The relationships between the tables that conform the database can be of three types: one-to-one, where each row in one table is linked to at most one row in another table; one-to-many, where a single row in Table A can be related to one or more rows in Table B but each row in table B is only related to one row in table A; and many-to-many, where multiple rows in one table can relate to multiple rows in another table, using a third table named junction or join table to manage the relationship between the two tables. However, these relationships are not mandatory and will not be enforced by the DBMS unless specified. The most common relationship type in relational databases is one-to-many. In them, there are two tables (A and B), each containing a column with identifier variables from the same domain, one with primary

keys, which uniquely identify each row within a table, and the other with foreign keys, which link one table to another by referencing a primary key. A foreign key is a column with the same primary keys as some table in the database. The relationship DBMS will use the relationship by matching data between primary and foreign keys to retrieve associated data, i.e., items from other columns. It is important to remark that relational data models place a constraint in one-to-may relationships, they require that each non-null foreign key value corresponds precisely to an existing primary key value. This is the most important constraint because it ensures the coherence of inter-table references.

To represent data relationship in a relational database we use entity-relationship (ER) diagram, which in practice is a diagram that shows relationship types among different tables (figure 1).

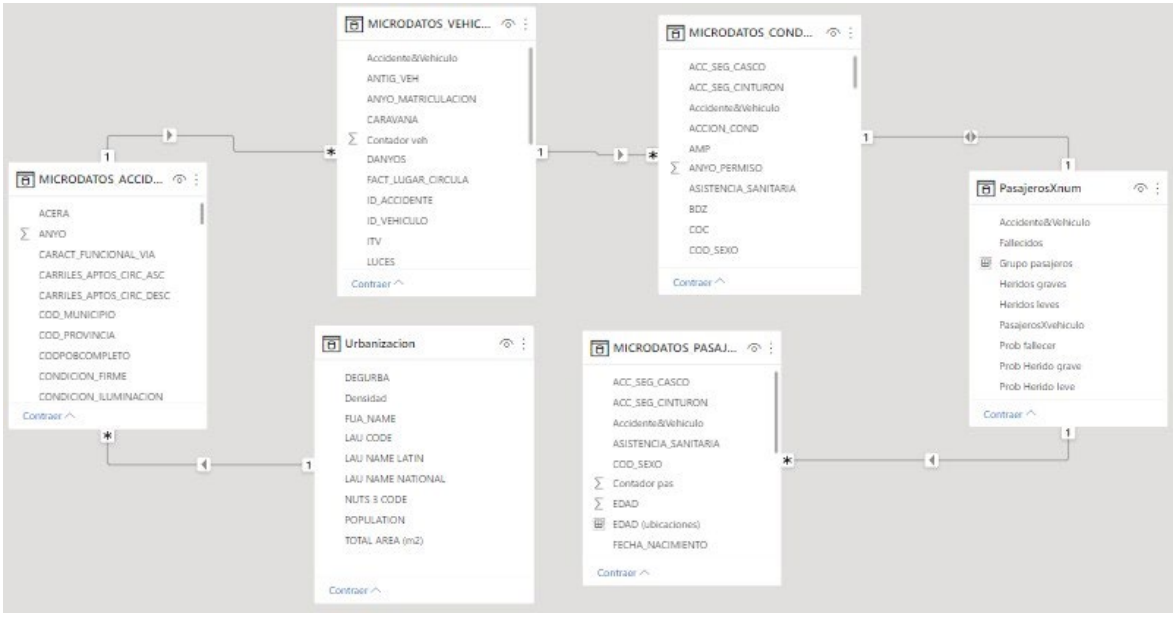

Figure 1. Entity-relationship diagram in Power BI. Source: Own elaboration. Note: in the figure, the different variables codes appear solely for illustrative purposes.

3. CARTOGRAPHIC LOCATION AND DEGREE OF URBANIZATION IN MOTOR INSURANCE

In European countries, according to Wijnen et al. (2019), the total costs of road crashes are equivalent to 0.4–4.1% of GDP. In the case of Spain, as the authors point out, the cost of road crashes stands at approximately 1% of the GDP. Moreover, Spanish insurance companies have seen their motor vehicle claims' soar in the last years, as shown by their net combined loss ratio (Table 1).

Table 1. Evolution of the Spanish Net Combined Loss Ratio for Motor Third-Party liability Q4-2020 to Q4-2023. Source: Own elaboration based on "Boletín de Información Trimestral de Seguros y Fondos de Pensiones Cuarto Trimestre 2023".

Year	2020	2021	2021	2021	2021	2022	2022
Quarter	4	1	2	3	4	1	2
Net Combined LR (%)	94	95.6	96.6	98.3	100.1	103.8	100.2

Year	2022	2022	2023	2023	2023	2023
Quarter	3	4	1	2	3	4
Net Combined LR (%)	100.6	103	104.6	105.7	105.1	107.1

The increases in their loss ratios underline the importance of understanding crashes better to reestablish profitability in the Motor Third Party Liability insurance. Incorporating additional spatial analysis could enhance their analytical process, allow for more geographical-based measures and contribute to the improvement of risk selection and pricing.

The riskiness of a driver is conditioned by the environment where he/she moves (Pljakić et al., 2022). Population density and its corresponding infrastructures not only heavily determines drivers' maneuvers and behaviors, but also the expected consequences of their mistakes (Abdel-Aty et al., 2013; Bil et al., 2019; Prato et al., 2018; Keeves et al., 2019). Several studies have shown that even if urban areas tend to concentrate a bigger proportion of crashes, occupants suffer worse injuries in rural areas crashes (Keeves et al., 2019, Peura et al., 2015). In Spain, the profile of the driver in rural areas is also different, older driver represent a bigger share of the census, according to the Spanish driver census for 2017-2019.

The intersection of rurality, rural depopulation and population ageing is important for Spain. At the same time, it is one of the European countries with the highest life expectancy, with the highest percentage of population living in cities, and the highest percentage of older people concentrated in rural areas (Gutiérrez et al., 2023; Casado-Sanz et al., 2019; European Commission, 2023). Rural municipalities represent 84% of Spain's surface area but only the 16% of the Spanish population lives in rural municipalities. The 25% of the rural population is over 65 years old and almost a third of those over 65 are over 80 years old. From the point of view of road safety, this context represents a huge challenge when it comes to ensuring safe and sustainable mobility in rural areas, which are increasingly depopulated and where the incidence of population aging is higher. In Spain, the greatest number of traffic accidents occur in urban environments but 52% of traffic fatalities occurs on rural environments (Harland et al., 2014).

Incorporating these structural demographic changes into insurance companies' models is important. However, adding new variables to a model is not trivial. It requires deciding how much resources to invest into making adjustments and ensuring accurate categorization. The difficulty increases with variables that do not have clear-cut classifications. Insurance companies are forced to choose between more variable customization or, when available, using standardized criteria. There are situations where opting for homogeneous criteria offers certain advantages, it makes market comparisons easier, and enables leveraging public available data if matched. In these situations, insurance companies could extract more insights from their policy holders and claims if they integrate their own (proprietary) data with information accessible from public sources.

An illustrative example is separating policyholders or claims by urbanization degree, i.e., whether it is urban or rural area. Although the use of geographical areas is very common in insurance, there is no harmonized widely-used criteria to determine the degree of urbanization of a location. Researchers and practitioners from different countries tend to use multiple measures (Harland et al., 2014), mostly related to population density or size, with different thresholds to determine urbanization and reflect their perspective on the urban – rural area dichotomy (Keeves et al., 2019). To ease international comparison and offer standardized classifications, Eurostat developed its own methodology (OECD, 2021). This European statistical institution defines cities, towns and rural areas based on a combination of population size, density and proximity and attributes a classification at the Local Administrative Unit (municipalities in the case of Spain). Using this methodology, it's possible to map the level of urbanization across EU member countries, enabling standardized comparison among them.

In Figure 2 we show results of applying the Eurostat's methodology in our country, including two maps at the municipal level. The first map indicates urbanization levels, while the second uses a logarithmic scale to show vehicle crash numbers involving injured victims in each municipality for the 2016-2019 period. Data of the numbers of crashes were provided by the Spanish General Traffic Directorate (DGT). The comparison suggests that a higher degree of urbanization correlates with an increase in traffic accidents. Mapping different factors can help to reveal potential relationship that provide insights into traffic crash dynamics, even helping

to predict crash occurrences based on certain criteria and to evaluate the potential correlation heterogeneity across different territories. For example, the impact of a factor on the number of crashes may vary, as it is the case in the maps of figure 2, where the rural areas (depicted in light green) correlate with a different magnitude in the northern and southern halves of Spain.

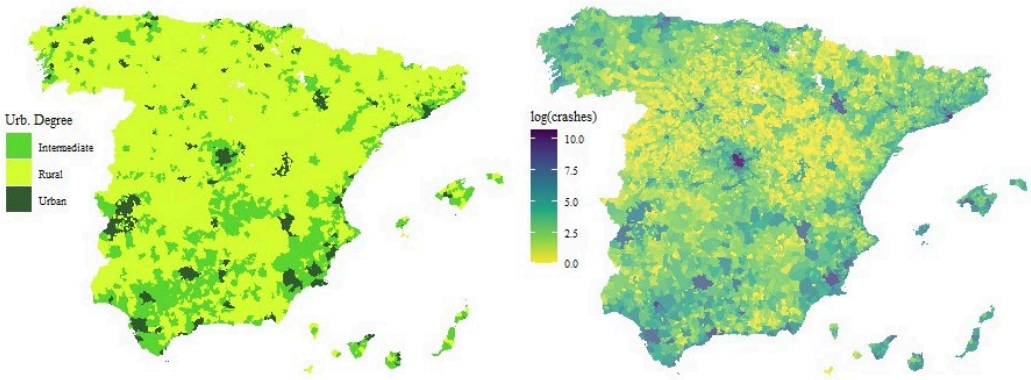

Figure 2. Map of Spanish peninsular municipalities by degree of urbanization in 2018 (right) and natural logarithm of the number of motor vehicle crashes 2016–2019 (left). Source: Cespedes et al. (2024b). Own elaboration.

The analysis of drivers' behavior patterns can also be enhanced by combining proprietary data with public information, such as that coming from Geographic Information Systems (GIS). The number of kilometers traveled influences the risk of accident (Boucher et al., 2013). Building from the segmentation into urbanization degrees, companies could extract the distances between the location of crashes and the residence of their policyholders, to study the differences according to the degree of urbanization, by querying distances between municipalities by means of the Open Source Routing Machine (OSRM). OSRM is an open-source routing engine that provides shortest routes in road network. An interface between R and OSRM is available by using the OSRM package (Giraud, 2022). Some patterns that emerge, for DGT data for the years 2016-2019, are that while the proportion of drivers crashing in their municipality of residence is equal to 63.9% in urban areas, it only is 26% in rural areas. If we consider drivers that crashed outside their municipality of residence and no further than 150 km from their homes, they represent a 30.1% of urban drivers, and 68.7% of rural drivers, with similar average distances (urban 32.4 km, rural 32.9 km) and standard deviations (urban 31.9 km, rural 29 km) (Cespedes et al., 2024a).

Using a simply relational data model that matches the postal code of the location of a claim or the residence of a driver to the degree of urbanization can give an edge to characterize more accurately these risks. Not only it allows companies to know the share of urban/suburban/rural drivers in the portfolio, to change their risk appetites and target desired proportion, but also to find commonalities in crashes or claims by urbanization degree, such that some of them can be bundled and analyzed together by pricing and reserving departments. Furthermore, other publicly available data could also be considered, such as the meteorology, as shown below, and enrich our understanding. These analyses open up a springboard to develop the skills needed to exploit telematics data in the future (Ayuso et al., 2014).

4. CLIMATOLOGICAL INFORMATION TO EXPLAIN INSURANCE CLAIMS IN MUNICIPALITIES

The climate of an area can be an important element in explaining the claims for insurance companies (Ashley et al., 2015; Eisenberg, 2004; Naik et al., 2016). Some examples in which weather may play a key role are motor insurance or home insurance, among others. Most of insurers already take meteorological information into account in their analysis of claim frequency and severity. In this section we show how insurers may incorporate public meteorological information in their claim analysis.

In this application we consider the annual number of accidents with victims per municipality in Spain for the year 2019. In total, there were 8,131 municipalities with at least 1 accident with victims. Figure 3 shows the number of accidents in the Spanish municipalities in the year 2019. Data are scaled per each thousand inhabitants to be comparable municipalities of different population size.

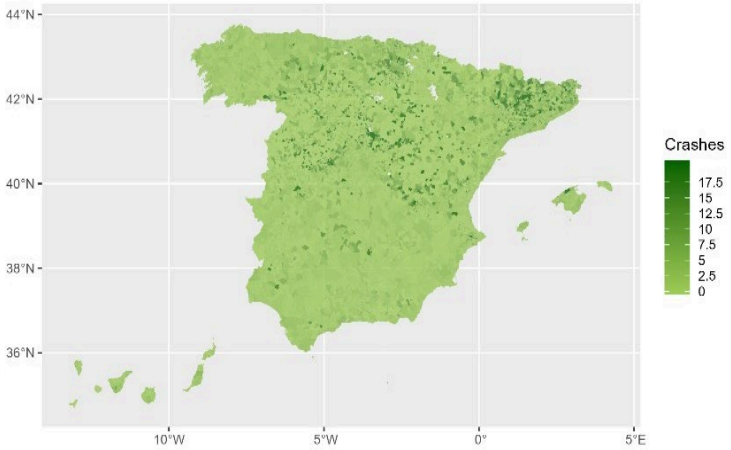

Figure 3. Number of motor crashes with victims per each 1,000 inhabitants in the Spanish municipalities in the year 2019. Source: Own elaboration.

Our objective in the example is to investigate whether weather information can be useful for insurers to explain the number of motor accidents with casualties. The accident data of the municipalities are adjusted to the number of inhabitants of the municipality. To avoid possible variability in the municipal motor accident rate due to the small population size of the municipality, we select municipalities with at least 500 inhabitants or more. The size of the dataset is now 4,146 Spanish municipalities with more than 500 inhabitants in which at least one motor crash with injured victims occurred in the year 2019. Table 3 shows descriptive statistics of the numerical variables of the dataset. Note that in this and the next sections we follow the definition of injury severity used by the DGT in which a victim is seriously injured if at least one day of hospitalization was required. Otherwise, injured victims are classified as casualties with slight injuries.[1] The degree of urbanization of the municipality following the European methodology used in the previous section is considered in the analysis. Table 4 shows the relative frequency of the degree of urbanization of the municipality (categorical variable with three categories).

[1] See Ayuso et al. (2019, 2020) for applications using the same bodily injury severity classification.

Table 3. Descriptive statistics for numerical variables in municipalities with more than 500 inhabitants. Source: Own elaboration based on DGT data (year 2019) and section 3.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Crashes	0	0.724	1.459	2.035	2.574	33.499
Involved vehicles	0	1.004	2.179	3.177	3.979	58.313
Casualties	0	0.906	1.971	2.962	3.702	52.109
Fatalities	0	0	0	0.083	0	7.005
Seriously injured casualties	0	0	0	0.291	0.297	16.304
Slight injured casualties	0	0.724	1.736	2.588	3.299	45.906
Average age of vehicles (in the municipality)	2.004	12.363	13.361	13.247	14.275	17.037
Population (in thousands)	0.501	1.021	2.334	11.276	6.861	3280.782
Percentage of male population	0.466	0.562	0.590	0.593	0.62	0.780

Table 4. Relative frequency of the degree of urbanization for municipalities with more than 500 inhabitants. Source: Own elaboration based on DGT data (year 2019) and section 3.

	Urban	Intermediate	Rural
Degree of urbanization	5.21	26.58	68.21

Now, we incorporate the climatic information of the municipalities in our dataset. We will use open data from the Spanish Agency of Meteorology (AEMET). In particular, we will use the normal climatological values corresponding to the period 1981-2010 for Spain for different climatic variables (AEMET, 2024).

The climatic variables of municipalities that can be consulted and that may be of interest to the insurance sector are:

- Average maximum daily rainfall (mm).
- Average annual and monthly accumulated precipitation (mm).
- Seasonal average accumulated precipitation (mm).
- Average annual number of days with precipitation greater than or equal to 0.1 mm.
- Average annual number of days with precipitation greater than or equal to 1 mm.
- Average annual number of days with precipitation greater than or equal to 10 mm.
- Average annual number of days with precipitation greater than or equal to 30 mm.
- Average annual and monthly temperature (ºC).
- Average annual and monthly minimum temperature (ºC).
- Average annual and monthly maximum temperature (ºC).
- Köppen-Geiger climate classification (Kottek et al., 2006).
- Mean annual number of snow days.
- Mean annual number of storm days.
- Mean annual number of fog days.
- Mean annual number of sunshine hours (Insolation).

The information available in AEMET is stored in GeoTIFF file format (.tif extension) that allows storing georeferenced information in an image file with TIFF format. Each GeoTIFF file corresponds to a raster image that refers to a climatological variable and period (monthly, annual or seasonal). The list of available files and the meteorological variables can be consulted in the appendix of Chazarra et al. (2018).

A raster image consists of a matrix of cells (pixels) organized in rows and columns in which each cell is represented by a color (Figure 4).

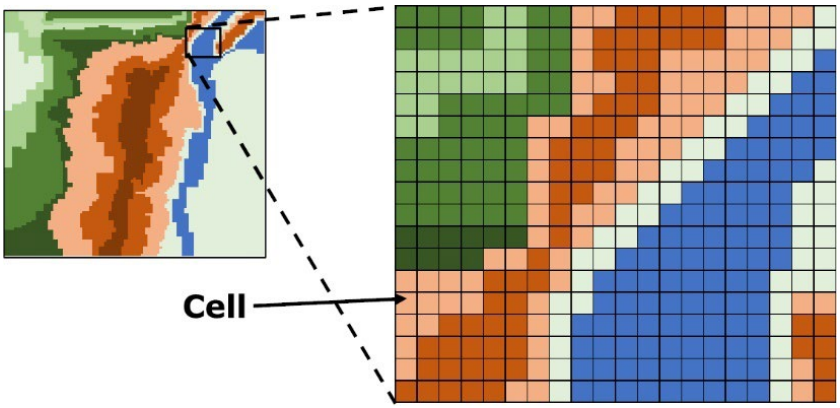

Figure 4. Example of a raster image. Source: own elaboration.

In the GeoTIFF files available at AEMET each cell is georeferenced. The images are projected according to the geographic coordinate system EPSG: 4326 (WGS 84 - WGS84 - World Geodetic System 1984). That is the most commonly used geographic coordinate system (used in Google Earth and GSP systems, for instance) and allows the geographic location of each cell of the image. The color of the cell represents the information of the value of the meteorological variable in that location. The colors of the image cells are defined in RGBA (Red, Green, Blue, Alpha) scale. Each parameter (Red, Green, and Blue) defines the intensity of the color between 0 and 255. The Alpha parameter represents the opacity/transparency, where 0 represents the maximum level of opacity (black), and 255 represents the maximum level of transparency. In our application, the alpha parameter would be particularly useful to distinguish between cells representing the land of the peninsula (or the island) and those that represent the sea. Finally, the GeoTIFF file provides scale information that links the RGBA colors of the cells with the values of the meteorological variable of interest.

To illustrate the use of meteorological information, we display the map of the average maximum daily rain precipitation in the 1981-2010 period in Canary Islands in Figure 5. The rainfall information was obtained from the GeoTIFF file downloaded from the AEMET website (AEMET, 2024). The equivalence between the RGBA space and the interval of the numerical values of the meteorological variable is provided at the bottom of the image, as a footnote.

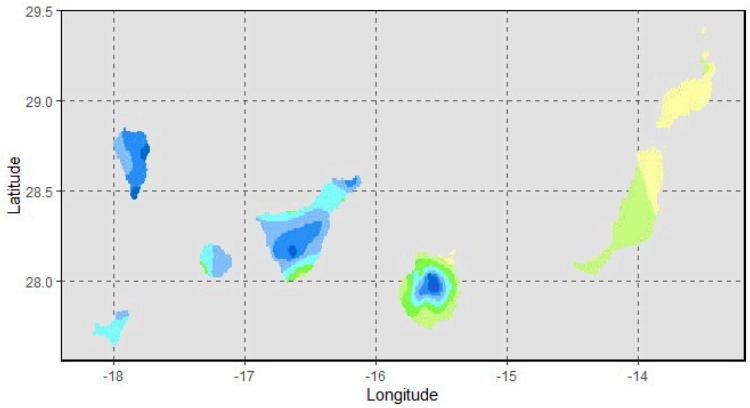

Figure 5. Average maximum daily rainfall (in mm) in the Canary Islands for the 1981-2010 period (in RGBA space). Source: Map of the AEMET. Scale: [{'Values': [140, ''], 'RGBA': ['255', '210', '255', '255']}, {'Values': [120, 140], 'RGBA': ['255', '138', '255', '255']}, {'Values': [100, 120], 'RGBA': ['162', '23', '253', '255']}, {'Values': [80, 100], 'RGBA': ['0', '106', '213', '255']}, {'Values': [70, 80], 'RGBA': ['41', '145', '248', '255']}, {'Values': [60, 70], 'RGBA': ['130', '191', '253', '255']}, {'Values': [50, 60], 'RGBA': ['128', '255', '255', '255']}, {'Values': [40, 50], 'RGBA': ['128', '255', '72', '255']}, {'Values': [30, 40], 'RGBA': ['201', '253', '130', '255']}, {'Values': [20, 30], 'RGBA': ['255', '255', '164', '255']}]}

Figure 5 was plotted downloading and reading the rainfall information from the raster image in RGBA space. The next step is to convert the RGBA information into a numerical value of the meteorological variable that insurers can incorporate in their insurance claim analysis. To achieve it, we previously convert the RGBA scale to the HSV scale. HSV space is a cylindrical-coordinate representation of points in an RGB color model. HSV stands for hue (type of color), saturation (quantity of color to be added), and value (brightness of the saturation of the color). Doing it, each color is represented by a single value and the numerical conversion of color values to the meteorological values becomes easier. Figure 6 represent the average maximum daily rainfall (in mm) in the Canary Islands for the 1981-2010 period after converting the color scale in a numerical value. The R code for downloading of the georeferencing raster image of the average maximum daily rainfall in Canary Islands and the steps to convert the information included in the GeoTIFF file into a numerical value of the meteorological variable of interest is provided in the Annex.

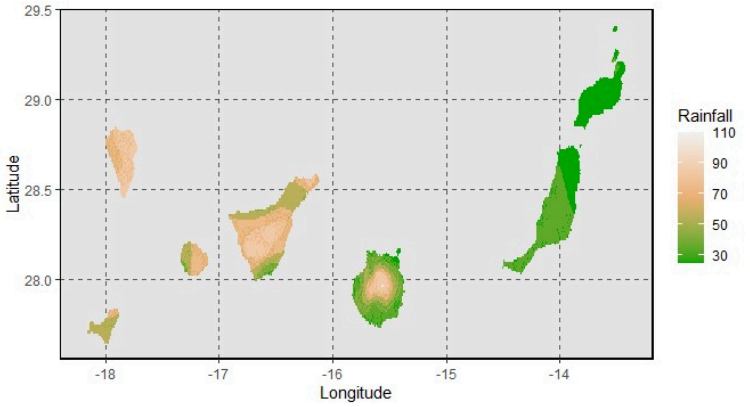

Figure 6. Average maximum daily rainfall (in mm) in the Canary Islands for the 1981-2010 period (in numerical value). Source: Own elaboration from GeoTIFF file of AEMET.

We carry out the same steps for the raster file containing rain precipitation information in the Peninsula and Balearic Islands. Finally, the numerical values of rain precipitations of referenced cells representing the Peninsula, the Balearic Islands and Canarian Islands are merged in a single file. Now, the information of rain precipitations can be incorporated into the database of motor crashes in Spanish municipalities with more than 500 inhabitants. We have longitude and latitude coordinates of municipalities, so we assign to the municipality the meteorological value of their longitude-latitude coordinates. Table 5 shows descriptive statistics for the variable of interest *average maximum daily rainfall (mm) in the period 1981-2010* for the municipalities with more than 500 inhabitants in which motor crashes involving victims occurred.

Table 5. Descriptive statistics of the meteorological variable *Average maximum daily rainfall (mm) in the period 1981-2010* for municipalities with more than 500 inhabitants. Source: Own elaboration.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Rain precipitations	25.00	45.00	55.00	59.58	75.00	150.00

For illustrative purposes, to demonstrate the explanatory capacity of the meteorological information on the number of motor crashes with casualties, a Poisson regression model has been fitted. We select as dependent variable the number of motor crashes involving victims in Spanish municipalities with more than 500 inhabitants. We include as regressors the average age of vehicles in the municipality where the accident took place, the percentage of male population in this municipality, the degree of urbanization of the municipality following the Eurostat methodology described in the previous section, and the average maximum daily rainfall (mm), as calculated in this section. Estimated coefficients are shown in Table 6. The variable associated with rain precipitations has an estimated positive coefficient statistically significant at 1% significance level. So, the amount of rain precipitations in a municipality seems to be positively related with the expected number of motor accidents involving victims in the municipality.

Table 6. Modeling the number of motor crashes involving victims in Spanish municipalities with more than 500 inhabitants (Generalized Linear Model-Poisson regression). Note: Population (in thousands) is included as offset of the regression model; Null deviance: 45,020; Residual deviance: 35,433.

Variable	Coef.	Std. Error	z value	p-value
Intercept	3.539	0.143	24.767	<0.001
Average age of vehicles (in the municipality)	-0.138	0.003	-52.043	<0.001
Percentage of male population	-3.533	0.283	-12.467	<0.001
Degree of urbanization – Rural	0.438	0.012	36.489	<0.001
Degree of urbanization – Urban	0.384	0.009	42.008	<0.001
Rain precipitations (mm)	0.006	<0.001	39.042	<0.001

Following with the example, a limitation of the use of Generalized Linear Models is that they are not flexible in the specification of the linear predictor (i.e., linear combination of parameters and regressors). In that sense, to allow for a nonlinear effect of the regressor associated with rain precipitations, we fit a Generalized Additive Model to investigate the functional form of the effect of rainfall on the (log) number of traffic accidents (Hastie, 1992). A Penalized spline (P −spline) is used to estimate the smooth function associated to the rainfall regressor (Eilers & Marx, 1996). The remaining regressors are included as a linear combination of parameters and regressors. Estimated coefficients of the linear predictor part are now shown in Table 7.

Table 7. Modelling the number of motor crashes involving victims in Spanish municipalities with more than 500 inhabitants with a P-spline for the rain precipitations (Generalized Additive Model-Poisson regression). Note: Population (in thousands) is included as offset of the regression model. R-sq.(adj) = 0.896; Deviance explained = 23.2%.

Variable	Coef.	Std. Error	z value	p-value
Intercept	3,982	0,148	26,835	<0.001
Average age of vehicles (in the municipality)	-0,133	0,003	-48,499	<0.001
Percentage of male population	-3,886	0,291	-13,351	<0.001
Degree of urbanization – Rural	0,433	0,012	35,825	<0.001
Degree of urbanization – Urban	0,352	0,009	38,029	<0.001

The estimated P-spline for the rain precipitations is shown in Figure 7. It can be observed that the effect of rainfall on the (log) number of accidents is increasing up to about the municipalities with average daily rainfall of 100 mm and then decreases. It is worth mentioning that only 3.7% municipalities took a value above 100 mm of average maximum daily rainfall. Even if the volume of traffic is reduced on wet days (Keay & Simmonds, 2005), the overall effect on crash rates depends on the increase in the relative risk of crash (Black et al., 2017). Therefore, the decreasing effect of rain precipitations in the right tail should be taken with caution and more analysis would be recommended.

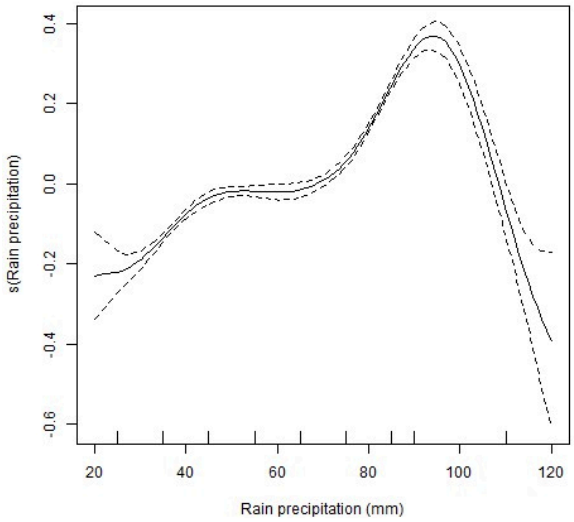

Figure 7. Estimated P-spline function for the rain precipitations to model the number of motor crashes with casualties in municipalities with more than 500 inhabitants (Generalized additive model-Poisson regression). Note: The effective degree of freedom of the P-spline (edf) is equal to 6.794.

5. BODILY INJURIES OF ALL OCCUPANTS OF THE CRASHED VEHICLE

Insurance companies have access to many data sources and can analyze motor claims from multiple perspectives. They can consider the characteristics of the location where the crash took place, as well as the injuries and characteristics of all occupants of the vehicles involved. For claims in which vehicles with passengers are involved, important information is lost when they only pay attention to aggregated costs and do not consider the variations coming from their injuries, especially when there is a recurrent pattern in the driver-passenger(s) profiles.

Given that crashes represent the subset of policies of the company's portfolio where the risk has materialized, crash reports are useful, not just to understand the claim, but also to infer traits of policyholders that condition their risk. For instance, understanding the heterogeneity in passenger injuries could help claims and reserving teams to have a more tailored opening reserves that do not distort in excess their quarterly average costs estimation.

Vehicles with passengers involved in injury crashes represent a relevant proportion of the total number of vehicles involved in injury crashes. In this section we use the dataset of motor crashes involving victims in Spanish roads in the period 2017-2019. According to the DGT, for the years 2017 to 2019, 19.6% of passenger cars that were involved in a crash in which at least one person was injured had passengers additional to the driver (Table 8). Given that the vehicle involved in the crash has at least one passenger, mostly has 1 passenger (65.7%), followed by 2 (20.4%) and 3 (10.1%) passengers. If attention is paid to their proportion of injuries according to their severity (Table 9), the proportion of serious and fatal injuries by number of passengers is very stable, around 2.1% and 0.5% correspondingly, while slight or no injuries change notably, as illustrated by the rejection of the equal distribution of injuries by passenger number in the Pearson's Chi-squared test. The more passengers there are in a car, the less likely they are to suffer slight injuries. This illustrates that it is important to consider the injury heterogeneity in the passenger vehicles to analyze the insured risk as a whole, both in terms of pricing and reserving. Understanding the characteristics of individuals with whom drivers tend to travel is a key issue, and from our knowledge, little treated in the automobile insurance context.

Table 8. Passenger frequency in motor vehicles. Source: Own elaboration from DGT databases 2017-2019.

Passengers (driver excluded)	Relative Frequency (%)	Relative Frequency ex 0 (%)
0	80.4	
1	12.9	65.7
2	4.0	20.4
3	2.0	10.1
4	0.7	3.5
5	0.1	0.3

Table 9. Proportion of injury types by number of passengers in passenger vehicles (%). Source: Own elaboration from DGT databases 2017-2019.

	Number of passengers (driver excluded)			
Injury Type	1	2	3	4
None	36.9	41.2	46.7	45.8
Slight	60.5	56.2	50.8	51.0
Serious	2.2	2.1	2.1	2.7
Fatal	0.5	0.4	0.5	0.5

Pearson's Chi-squared test P-value < $2.2 * e^{-16}$

Rather than solely evaluating the total cost per vehicle for the claim, a more nuanced analysis can be conducted. It involves estimating the conditional expected cost of the claim based on both the driver's characteristics and those of its expected passengers. For instance, if a crash occurred knowing that the vehicle had only one passenger alongside the driver, an initial reserve for bodily injury could be established based on the expected probabilities of severity levels according to the number of passengers (Table 9). Alternatively, if the company wanted to estimate the total expected costs of bodily injury claims, it could incorporate the expected probability that insured drivers will be accompanied by one or more passengers (Table 8) and the estimated severity of injuries in each case (Table 9). Matching all the tables can be easily done with a relational database, where keys are matched to relate passengers or all occupants with crash characteristics, enabling data manipulation and analysis. The composition inside

the car varies and multiple combinations may be considered. For instance, the analysis could be made by gender of driver and passengers, by different age groups, or by geographical area, linking with research presented at previous sections.

The diversity of passenger profiles is easily appreciated in the data, as illustrated for example by Tables 10 and 11, based on DGT data. In Table 10, the relative frequency of the gender of passenger by injury type and gender of the driver can be seen; e.g., of all passengers with fatal injuries and a male driver, 46% were men, while the remaining 54% were women. In this table, it can be seen that women tend to have relatively better outcomes with a woman driver except for fatalities, where they represent a smaller proportion of all fatalities. Statistical independence between the driver-passenger gender pairing and the passenger's injury type using the Pearson's chi-square test for independence returned a p-value <0.01.

Table 10. Passenger injury by driver gender (%). Source: Own elaboration from DGT databases 2017-2019.

Gender		Passenger Injury			
Driver	Passenger	None	Slight	Serious	Fatal
Man	Man	44.1	35.8	40.5	46.0
Man	Woman	55.9	64.2	59.5	54.0
Gender		Passenger Injury			
Driver	Passenger	None	Slight	Serious	Fatal
Woman	Man	49.9	38.3	43.6	36.9
Woman	Woman	50.1	61.7	56.4	63.1

In Table 11, the relative frequency of the pairing of a driver and passenger by age and number of passengers in the crashed vehicles can be seen. Statistical independence between the driver-passenger age pairing and the passenger's number using the Pearson's chi-square test for independence returned a p-value <0.01. For vehicles with 1 passenger, 87% of them were driven by a person aged 18-64 years old, while the remaining 13% were accompanied by a driver older than 64 years old. Also, the more passengers there are in the vehicle, the bigger the proportion of drivers aged 18-64 years old: 93.9% for 2 passengers, 94.6% for 3, and 96% for vehicles with 4 passengers. Therefore, the first suggestion an observer would consider is that most passengers, regardless of age, go on the road with drivers younger than 65 years old. However, if passengers are grouped by age and the relative frequency by driver age is considered, as presented in Table 12, we get a different result. Older adult passengers go on the road with drivers older than 64 years old in a significant proportion. In 68.5% of vehicles with 1 passenger and a passenger aged 65 or more years, the driver is also and older adult. In vehicles with 2,3, and 4 passengers, if there is an older adult passenger, the likelihood of having an older driver are 35.6%, 43.3%, and 29.2% correspondingly.

Table 11. Relative frequencies (%) of the passenger - driver age by number of passengers. Source: Own elaboration from DGT databases 2017-2019.

		Passengers			
Driver age	Passenger age	1	2	3	4
[18, 65)	[18, 65)	82.9	89.2	91.2	92.6
[18, 65)	[65,)	4.1	4.7	3.4	3.4
[65,)	[18, 65)	4.1	3.5	2.8	2.6
[65,)	[65,)	8.9	2.6	2.6	1.4
	Total	100%	100%	100%	100%

Table 12. Relative frequencies (%) of the passenger - driver age by number of passengers and driver age. Source: Own elaboration from DGT databases 2017-2019.

		Passengers			
Driver age	Passenger age	1	2	3	4
[18, 65)	[18, 65)	95.3%	96.2%	97.0%	97.3%
[65,)	[18, 65)	4.7%	3.8%	3.0%	2.7%
	Total	100%	100%	100%	100%
[18, 65)	[65,)	31.5%	64.4%	56.7%	70.8%
[65,)	[65,)	68.5%	35.6%	43.3%	29.2%
	Total	100%	100%	100%	100%

To exemplify how to combine the different databases we are using in this paper, we select now passenger cars with 5 or less occupants. The insurer knows some drivers' characteristics, such as residence and gender. If the Eurostat methodology is applied, we obtain, in Table 13, the following relative frequency of driver gender-residence-age combinations.

Table 13. Relative Frequencies (%) of the gender, residence degree of urbanization, and age of driver in this segmentation. Source: Own elaboration.

Driver Gender	Residence Urb. Degree	Age group	Rel. Freq. (%)
Men	Rural	[18, 65)	10.1%
Men	Rural	[65,)	2.0%
Men	Urban	[18, 65)	46.2%
Men	Urban	[65,)	6.6%
Women	Rural	[18, 65)	6.0%
Women	Rural	[65,)	0.3%
Women	Urban	[18, 65)	27.4%
Women	Urban	[65,)	1.3%

Now, we consider the proportion of occupants by each combination. It can be seen that proportions for the different groups seem to be different (Table 14).

Table 14. Proportion of vehicles with a determined number of occupants by driver with a combination of gender, residence urbanization degree, and age group. Source: Own elaboration.

			Number of vehicle's occupants					
Driver Gender	Residence Urb. Degree	Age group	1	2	3	4	5	Total
Men	Rural	[18,65)	67.9%	20.7%	6.6%	3.5%	1.2%	100%
Men	Rural	[65,)	68.4%	26.9%	3.3%	1.1%	0.3%	100%
Men	Urban	[18,65)	72.5%	17.8%	5.8%	3.0%	0.9%	100%
Men	Urban	[65,)	68.3%	26.1%	3.6%	1.6%	0.3%	100%
Women	Rural	[18,65)	72.2%	18.3%	6.4%	2.2%	0.8%	100%
Women	Rural	[65,)	78.0%	18.2%	3.1%	0.4%	0.2%	100%
Women	Urban	[18,65)	77.6%	14.9%	4.9%	2.0%	0.6%	100%
Women	Urban	[65,)	79.0%	16.2%	3.1%	1.4%	0.3%	100%
		Total	73.2%	18.1%	5.4%	2.6%	0.8%	100%

The possibilities that open up in terms of segmenting risks and optimizing pricing and reservation processes are evident. Thanks to relational models and the combination of databases, risk characterization can now be done with a much higher level of detail.

6. CONCLUSION

The access to data has exponentially grown in last years and more diverse data sources become available to insurers. Techniques to link different databases to extract useful information become more important in pricing and risk management. As a previous step to delving into complex techniques, it is crucial to explore the interrelation of data to understand better the behaviour of our policyholders. We have observed there are cost-effective uses of resources already available to insurance companies, such as relational databases, but to leverage them, they must be creative and experiment to find useful relationships for them. Relational databases serve as a valuable tool in exploring and extending it to all areas of insurance companies.

There are potential applications all over the industry, combining in-house data with public available data can give them an edge, as evidenced in this paper by using spatial analysis and attributing a standard urbanization degree, or considering climatological variables such as rain. In both cases, we were able to add variables that helped us to grasp better the heterogeneity in our data, allowing for more adjusted general analysis and the possibility to split the database into more homogeneous segmentations. Also, the use of standardized categorizations can help companies to compare themselves against industry benchmarks, making comparisons faster and easier, by avoiding arbitrary specifications. Let us highlight the relevance of these analyzes in the new Sustainability framework, where factors such as climate change or population relocation take on a leading role.

Nevertheless, relational databases with only in-house data can be advantageous too, as long as companies are creative and curious, as illustrated by the study of the passengers' heterogeneity of the same vehicle. In this study we use motor crash data from the DGT but similar analysis can be done by insurers with in-house data. Depending on the needs of the company, the depth of analysis can vary, but we have seen that there is room for conditional analysis, that some important variables have known-values beforehand, such as gender, age or residence urbanization degree, that can help to forecast crashes' outcomes and estimate its variability based on past experience.

7. ACKNOWLEDGMENTS

We are grateful to the Dirección General de Tráfico for access to their database. We also are grateful to the Spanish Ministry of Science and Innovation grant PID2019-105986GB-C21 and to the Departament de Recerca i Universitats, del Departament d'Acció Climàtica, Alimentació i Agenda Rural i del Fons Climàtic de la Generalitat de Catalunya (2023 CLIMA 00012).

8. REFERENCES

Abdel-Aty, M., Lee, J., Siddiqui, C., & Choi, K. (2013). Geographical unit-based analysis in the context of transportation safety planning. Transportation Research. Part A, *Policy and Practice*, 49, 62–75

AEMET (2024). Valores climatológicos normales, Agencia Estatal de Meteorología. https://www.aemet.es/es/serviciosclimaticos/datosclimatologicos/valoresclimatologicos. Accessed June 12 2024

Alarifi, S. A., Abdel-Aty, M. A., Lee, J., & Park, J. (2017). Crash modeling for intersections and segments along corridors: A Bayesian multilevel joint model with random parameters. *Analytic Methods in Accident Research*, 16, 48–59. https://doi.org/10.1016/j.amar.2017.08.002

Ashley, W. S., Strader, S., Dziubla, D. C., & Haberlie, A. (2015). Driving blind: Weather-related vision hazards and fatal motor vehicle crashes. *Bulletin of the American Meteorological Society*, 96(5), 755-778

Ayuso, M., Guillén, M., & Pérez-Marín, M. (2014) Los hábitos de conducción al volante según el género en los seguros pay-as-you-drive o usage-based. *Anales del Instituto de Actuarios Españoles*, 20, 17-32

Ayuso, M., Sanchez, R., & Santolino, M. (2019). Longevidad de los conductores y antigüedad de los vehículos: impacto en la severidad de los accidentes. *Anales del Instituto de Actuarios Españoles*, 25, 33-53

Ayuso, M., Sanchez, R., & Santolino, M. (2020). Does longevity impact the severity of traffic crashes? A comparative study of young-older and old-older drivers. *Journal of Safety Research*, 73, 37-46

Bil, M., Andrasik, R., & Sedonik, J. (2019). Which curves are dangerous? A network-wide analysis of traffic crash and infrastructure data. Transportation Research. Part A, *Policy and Practice*, 120, 252–260

Black, A. W., Villarini, G., & Mote, T. L. (2017). Effects of Rainfall on Vehicle Crashes in Six U.S. States. *Weather, Climate, and Society*, 9(1), 53–70. https://doi.org/10.1175/WCAS-D-16-0035.1

Boucher, J. P., Pérez-Marín, A. M., & Santolino, M. (2013). Pay-As-You-Drive insurance: the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles*, 3ª época, 19, 135-154

Casado-Sanz, N., Guirao, B., & Gálvez-Pérez, D. (2019). Population ageing and rural road accidents: Analysis of accident severity in traffic crashes with older pedestrians on Spanish crosstown roads. *Research in Transportation Business & Management*, 30, 100377. https://doi.org/10.1016/j.rtbm.2019.100377

Cespedes, L. E., Ayuso, M., & Santolino, M. (2024a). *Distance between the driver's residence and the motor accident: is an insurance risk factor?* UB RISKcenter Working paper, in progress

Cespedes, L. E., Santolino, M., & Ayuso, M. (2024b). *Population density in aging societies and severity of motor vehicle crash injuries: the case of Spain*. (Submitted)

Chazarra Bernabé, A., Flórez García, E., Peraza Sánchez, B., Tohá Rebull, T., Lorenzo Mariño, B., Criado Pinto, E., Moreno García, J.V., Romero Fresneda, R., & Botey Fullat, R. (2018). *Mapas climáticos de España (1981-2010) y ETo (1996-2016)*. Ministerio para la Transición Ecológica, Agencia Estatal de Meteorología, Madrid

Clark, D. E., & Cushing, B. M. (2004). Rural and urban traffic fatalities, vehicle miles, and population density. *Accident Analysis and Prevention*, 36(6), 967–972. https://doi.org/10.1016/j.aap.2003.10.006

Codd, E. F. (1970). A relational model of data for large shared data banks. In *Communications of the ACM*, 13(6), 377-387

Denuit, M., Hainaut, D., & Trufin, J. (2020). Effective statistical learning methods for actuaries II : tree-based methods and extensions (1st ed. 2020). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-57556-4

Eilers, P. H. C., & Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89-121. https://doi.org/10.1214/ss/1038425655

Eisenberg, D. (2004). The mixed effects of precipitation on traffic crashes. *Accident Analysis and Prevention*, 36(4), 637–647. https://doi.org/10.1016/S0001-4575(03)00085-X

European Commission (2023). *Road safety in the EU: fatalities below pre-pandemic levels but progress remains too slow*. Available at: https://transport.ec.europa.eu/news-events/news/road-safety-eu-fatalities-below-pre-pandemic-levels-progress-remains-too-slow-2023-02-21_en. Accessed 29 February 2024

Giraud, T. (2022). osrm: Interface Between R and the OpenStreetMap-Based Routing Service OSRM. *Journal of Open Source Software*, 7(78), 4574. doi:10.21105/joss.04574, https://doi.org/10.21105/joss.04574

Gutiérrez, E., Moral-Benito, E., Oto-Peralías, D., & Ramos, R. (2023). The spatial distribution of population in Spain: An anomaly in European perspective. *Journal of Regional Science*, 63(3), 728–750. https://doi.org/10.1111/jors.12638

Harland, K. K., Greenan, M., & Ramirez, M. (2014). Not just a rural occurrence: Differences in agricultural equipment crash characteristics by rural–urban crash site and proximity to town. *Accident Analysis and Prevention*, 70, 8–13. https://doi.org/10.1016/j.aap.2014.02.013

Harrington, J. L. (2016). *Relational database design and implementation* (Fourth edition). Morgan Kaufmann, an imprint of Elsevier

Hastie, T. J. (1992). *Generalized Additive Models. Statistical Models in S* (1st ed.). Routledge. https://doi.org/10.1201/9780203738535

Haynes, R., Lake, I. R., Kingham, S., Sabel, C. E., Pearce, J., & Barnett, R. (2008). The influence of road curvature on fatal crashes in New Zealand. *Accident Analysis and Prevention*, 40(3), 843–850. https://doi.org/10.1016/j.aap.2007.09.013

Keay, K., & Simmonds, I. (2005). The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accident Analysis and Prevention*, 37(1), 109–124. https://doi.org/10.1016/j.aap.2004.07.005

Keeves, J., Ekegren, C. L., Beck, B., & Gabbe, B. J. (2019). The relationship between geographic location and outcomes following injury: A scoping review. *Injury*, 50(11), 1826–1838. https://doi.org/10.1016/j.injury.2019.07.013

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, Vol. 15, No. 3, 259-263

Lee, J., Abdel-Aty, M., & Choi, K. (2014). Analysis of residence characteristics of at-fault drivers in traffic crashes. *Safety Science*, 68, 6–13. https://doi.org/10.1016/j.ssci.2014.02.019

Lerner, E. B., Jehle, D. V. K., Billittier, A. J., Moscati, R. M., Connery, C. M., & Stiller, G. (2001). The influence of demographic factors on seatbelt use by adults injured in motor vehicle crashes. *Accident Analysis and Prevention*, 33(5), 659–662. https://doi.org/10.1016/S0001-4575(00)00080-4

Liu, C., & Sharma, A. (2018). Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic Methods in Accident Research*, 17, 14–31. https://doi.org/10.1016/j.amar.2018.02.001

Lord, D., Manar, A., & Vizioli, A. (2005). Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis and Prevention*, 37(1), 185–199. https://doi.org/10.1016/j.aap.2004.07.003

Moeinaddini, M., Asadi-Shekari, Z., & Zaly Shah, M. (2014). The relationship between urban street networks and the number of transport fatalities at the city level. *Safety Science*, 62, 114–120. https://doi.org/10.1016/j.ssci.2013.08.015

Naik, B., Tung, L. W., Zhao, S., & Khattak, A. J. (2016). Weather impacts on single-vehicle truck crash injury severity. *Journal of Safety Research*, 58, 57–65. https://doi.org/10.1016/j.jsr.2016.06.005

OECD (2021). *Applying the Degree of Urbanisation A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons*. 2021 edn.OECD Publishing. Paris.

Peura, C., Kilch, J. A., & Clark, D. E. (2015). Evaluating adverse rural crash outcomes using the NHTSA State Data System. *Accident Analysis and Prevention*, 82, 257–262. https://doi.org/10.1016/j.aap.2015.06.005

Pljakić, M., Jovanović, D., & Matović, B. (2022). The influence of traffic-infrastructure factors on pedestrian accidents at the macro-level: The geographically weighted regression approach. *Journal of Safety Research*, 83, 248–259. https://doi.org/10.1016/j.jsr.2022.08.021

Prato, C. G., Kaplan, S., Patrier, A., & Rasmussen, T. K. (2018). Considering built environment and spatial correlation in modelling pedestrian injury severity. *Traffic Injury Prevention*, 19(1), 88-93 (2018). https://doi.org/10.1080/15389588.2017.1329535

R Core Team (2024). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Wang, C., Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety Science*, 57, 264–275. https://doi.org/10.1016/j.ssci.2013.02.012

Watt, A., & Eng, N. (2014). *Database Design*. 2nd edn. BCcampus, British Columbia

Wijnen, W., Weijermars, W., Schoeters, A., van den Berghe, W., Bauer, R., Carnis, L., Elvik, R., & Martensen, H. (2019). An analysis of official road crash cost estimates in European countries. *Safety Science*, 113, 318–327. https://doi.org/10.1016/j.ssci.2018.12.004

Wüthrich, M. V., & Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications* (1st ed. 2023.). Springer Nature. https://doi.org/10.1007/978-3-031-12409-9

Xie, K., Yang, D., Ozbay, K., & Yang, H. (2019). Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure. *Accident Analysis and Prevention*, 125, 311–319. https://doi.org/10.1016/j.aap.2018.07.002

Zheng, L., Sayed, T., & Essa, M. (2019). Validating the bivariate extreme value modeling approach for road safety estimation with different traffic conflict indicators. *Accident Analysis and Prevention*, 123, 314–323. https://doi.org/10.1016/j.aap.2018.12.007

Zheng, L., Sayed, T., & Mannering, F. (2021). Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. *Analytic Methods in Accident Research*, 29, 100142-. https://doi.org/10.1016/j.amar.2020.100142

Ziakopoulos, A. (2024). Analysis of harsh braking and harsh acceleration occurrence via explainable imbalanced machine learning using high-resolution smartphone telematics and traffic data. *Accident Analysis and Prevention*, 207, 107743-. https://doi.org/10.1016/j.aap.2024.107743

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis and Prevention*, 135, 105323–105323. https://doi.org/10.1016/j.aap.2019.105323

9. APPENDIX

R Software was used in the applications (R Core Team, 2024). R code used to plot Figures 3 and 4 is provided below.

```
# Download georeferenced data.
# AEMET (2024), Normal climatological values, Agencia Estatal de Meteorología
https://www.aemet.es/es/serviciosclimaticos/datosclimatologicos/valoresclimatologicos [Access: June 12, 2024].

# Description: download of all climatic variables for Peninsula and Balearic Islands, and Canary Islands is available.
# SGR: EPSG:4326 (WGS84 - World Geodetic System 1984).
# Download unit: each raster image corresponds to a variable and period (monthly, annual or seasonal).
# Format: GeoTIFF (.tif)
# Additional information: the "SCALE" field provides the correspondence between the RGBA color scale and the range of values of the meteorological variable.

# Chazarra Bernabé, A., Flórez García, E., Peraza Sánchez, B., Tohá Rebull, T.,  Lorenzo Mariño, B., Criado Pinto, E., Moreno García, J.V., Romero Fresneda, R. y Botey Fullat, R. (2018) Mapas climáticos de España (1981-2010) y ETo (1996-2016) Ministerio para la Transición Ecológica, Agencia Estatal de Meteorología, Madrid
# Allows to identify file name with climatological variable and image of peninsula or canary islands.
```

```
getRversion()

## [1] '4.3.0'

# Packages
library(raster) #read tif file

library(rgdal) #from source file (tar.gz). Used to obtain scale information

library(ggplot2)

library(grDevices) #convert RGB (red/green/blue) values in HSV (hue/saturation/value).
```

To read the raster file and consult contained information.

```
# Working directory
setwd("~\\anales\\clima\\descarga_clima")

# Rain precipitations in Canary Islands

# Information of the scale and other infromation
rgdal::GDALinfo("down_vn8110pxdmww13c_20170512.tif")

Canarias <- stack("down_vn8110pxdmww13c_20170512.tif") #create a multi-layer raster object
df <- as.data.frame(Canarias, xy= TRUE) #data frame includes longitude, latitude and RGBA information
names(df)<-c("long","lat","R","G","B","A") #change name
```

To plot Figure 5.

```
dfclean<-df[df$A>100,] #values less than 100 are removed (cells indicating the sea)

Fig3<-ggplot(data = dfclean, aes(x = long, y =lat))+
  geom_raster(aes(fill=rgb(R,G,B, maxColorValue = 255))) +
  scale_fill_identity()+xlab("Longitude")+ylab("Latitude")
```

To convert RGBA values in numerical values of the meteorological variable of interest.

```
# Scale in RGBA space

escala<-matrix(c(
255, 210, 255, 255,
255, 138, 255, 255,
162, 23, 253, 255,
0, 106, 213, 255,
41, 145, 248, 255,
130, 191, 253, 255,
128, 255, 255, 255,
128, 255, 72, 255,
201, 253, 130, 255,
255, 255, 164, 255), ncol=4, byrow=T)

# Convert to HSV space
escalhsv<-t(rgb2hsv(r=escala[,1],g=escala[,2],b=escala[,3]))

# Function that assigns discrete numerical value of the climatic variable to the HSV color value.

Meteo<-function(A){
C<-t(rgb2hsv(A[,1], A[,2] , A[,3]))
valor<-cbind(C,NA)
colnames(valor)[4]<-"val"

valor<-as.matrix(as.data.frame(valor))
cond1<-I(valor[,1]<=0.10)
valor[(cond1==T),4]<-1

cond2<-I(valor[,1]>0.10 & valor[,1]<=0.22)
valor[(cond2==T),4]<-25

cond3<-I(valor[,1]>0.22 & valor[,1]<=0.28)
valor[(cond3==T),4]<-35

cond4<-I(valor[,1]>0.28 & valor[,1]<=0.49)
valor[(cond4==T),4]<-45

cond5<-I(valor[,1]>0.49 & valor[,1]<=0.55)
valor[(cond5==T),4]<-55

cond6<-I(valor[,1]>0.55 & valor[,1]<=0.65)
cond6A<-I(valor[,2]>0.83)
cond6B<-I(valor[,2]>0.48 & valor[,2]<=0.83)
cond6C<-I(valor[,2]<=0.48)
valor[(cond6==T)&(cond6A==T) ,4]<-85
valor[(cond6==T)&(cond6B==T),4]<-75
valor[(cond6==T)&(cond6C==T),4]<-65

cond7<-I(valor[,1]>0.65 & valor[,1]<=0.76)
valor[(cond7==T),4]<-90

cond8<-I(valor[,1]>0.76 & valor[,1]<=0.83)
valor[(cond8==T),4]<-110

cond9<-I(valor[,1]>0.83 & valor[,1]<=1)
cond9A<-I(valor[,2]<=45)
cond9B<-I(valor[,2]>45)
valor[(cond9==T)&(cond9A==T),4]<-150
valor[(cond9==T)&(cond9B==T),4]<-130
valor[valor[,3]<0.4,4]<-NA  # Values of v close to zero (black) are considered missing
return(valor)
}
```

To plot Figure 6.

```
A<-cbind(dfclean[, 3:5]) # matrix with color values
lluvia<-Meteo(A)

dfclean$Lluvia<-lluvia[,4] # numerical value

Fig4<- ggplot(data = dfclean, aes(x = long, y =lat))+
  geom_raster(aes(fill=Lluvia), show.legend = T) +
  scale_fill_identity()+xlab("Longitude")+ylab("Latitude")+ scale_fill_gradientn(colours =
terrain.colors(3))+ labs(fill = "Rainfall")
```