



Dataset of an actual motor vehicle insurance portfolio

Jorge Segura-Gisbert¹ · Josep Lledó¹ · Jose M. Pavía¹

Received: 28 May 2024 / Revised: 30 July 2024 / Accepted: 15 August 2024
© The Author(s) 2024

Abstract

Advanced analytics plays a vital role in enhancing various aspects of business operations within the insurance sector by providing valuable insights that drive informed decision-making, primarily through effective database utilization. However, open access databases in the insurance industry are exceedingly rare, as they are the basis of the business, encapsulating all the risk structure of the company. This makes it challenging for researchers and practitioners to access comprehensive insurance datasets for analysis and assessing new approaches. This paper introduces an extensive database specifically tailored for non-life motor insurance, containing 105,555 rows and encompassing a wide array of 30 variables. The dataset comprises important date-related information, such as effective date, date of birth of the insured, and renewal date, essential for policy management and risk assessment. Additionally, it includes relevant economic variables, such as premiums and claim costs, for assessments of products' financial profitability. Moreover, the database features an array of risk-related variables, such as vehicle size, economic value, power, and weight, which significantly contribute to understanding risk dynamics. By leveraging the statistical analysis of this rich database, researchers could identify novel risk profiles, reveal variables that influence insured claims behaviour, and contribute to the advancement of educational and research initiatives in the dynamic fields of economics and actuarial sciences. The availability of this comprehensive database opens new opportunities for research and teaching and empowers insurance professionals to enhance their risk assessment and decision-making processes.

Keywords Non-life insurance · Premium · Claims · Profitability

✉ Josep Lledó
josep.lledo@uv.es

¹ Department of Applied Economics, Quantitative Methods Area, Universitat de Valencia, Valencia, Spain

1 Introduction

Open-access databases in the insurance industry are exceedingly rare. Companies are typically reluctant to share their data as it constitutes the core of their business. During a research project conducted by the authors within a Spanish insurance company, specialized in non-life insurance, we gained access to one of their motor vehicle insurance portfolio datasets and were authorized to share a sample of anonymized data. The data is presented in the form of a structured table, providing valuable insights into a diverse range of non-life insurance policies. Finding these data valuable for our own research, we believe they can be also of interest to other researchers and educators. This data can be used to identify novel risk profiles, reveal variables that influence insured claims behaviour, and contribute to training future actuaries.

This paper presents and describes the database designed with a collection of 105,555 rows of data and encompassing a diverse set of 30 variables. Within this dataset, one can find indispensable date-related information, including the effective date of policies, the birthdates of insured individuals, and renewal dates. Furthermore, this dataset is enriched with pivotal economic variables, notably premiums and claim costs. These economic indicators are necessary for conducting in-depth analyses of the financial viability and profitability of motor insurance products.

These data can be used in several areas of an insurance company, being relevant for marketing purposes; including customer segmentation, contract renewal processes, price renewal strategies, optimization and price sensitivity models, as well as pricing mechanisms for new business. Researches can use these data to test risk selection mechanisms and the utilization of risk underwriting standards in the insurance sector, also in the context of Solvency II framework. Specifically, it could be employed to examine how insurance companies determine the acceptance or conditional acceptance, as well as the rejection, of risks by adapting their own established underwriting standards. The database can be utilized not only in the professional and research fields but also in the educational sector. It can be applied to non-life insurance subjects, allowing students to practice with real data the theoretical techniques learned in class.

The sources of the data originate from a non-life insurance company operating in Spain; therefore, the data have been meticulously anonymized to align with the prevailing European legislation, safeguarding individual privacy and confidentiality. For ease of access and reference, a supplementary clean data file in spreadsheet format is provided along with this article. Researchers and stakeholders interested in exploring and using these data can access them via the provided DOI link: <https://doi.org/10.17632/5cxyb5fp4f.2>, or by contacting the authors via email. The pre-cleaned raw data can also be obtained from the authors.

2 Background

The insurance industry plays an indispensable role in driving and supporting modern economies. At its core, the purpose of insurance is to safeguard agents and individuals against adverse events that may lead to economic losses or repercussions.

This crucial function is realized through insurance policies, which aim to provide compensation or indemnity in the event of a loss.

In the digital age, the insurance sector is undergoing transformative changes, and a key driving force behind this transformation is a more intensive utilization of data. Entities within the industry are leveraging data from various sources, both internal and external, to improve their decision-making processes. By combining diverse information assets, insurers gain a competitive advantage, fostering innovation and sector-wide transformation.

Notably, customer behaviour and risk aversion vary among individuals in the insurance industry. Therefore, analysing the profitability of customers and policyholders becomes crucial for shaping and adapting business strategies. Insurers are wary of consistently facing losses from policyholders, making it essential to identify and engage with customers who contribute positively to the bottom line. In this sense, Venkatesan and Kumar [7] highlight that the true measure of success of a business lies not in the sheer number of customers in a portfolio, but in the quality of the customers and the profitability of the relationships established.

The availability of open access data concerning insured populations is currently limited. As exceptions we can list Lledó and Pavía [2], New York State Department of Financial Services [3–5], and Department of Health and Human Services [1]. In life insurance, Lledó and Pavía [2] offers a large database of a real portfolio containing valuable information concerning, among other issues, ages of insured with capital at risks and quarterly identifiers. In the context of non-life insurance, one can find databases pertain to particular areas, such as car insurance, property insurance, and health insurance. Regarding car insurance, the database New York State Department of Financial Services [3] contains valuable information based on upheld consumer complaints as a percentage of their total business over two years. Insurers with the lowest complaint ratios rank higher on the list, while those with the highest ratios are placed at the bottom. New York State Department of Financial Services [4] determines the number of motorists who have completed the course and organize the data into subsets based on the year and month of course completion, motorist residency, age, and sex, as well as the course provider and delivery method. Similarly, New York State Department of Financial Services [3] and Department of Health and Human Services [1] are two datasets for property and health insurance sector respectively. Nevertheless, none of the existing databases matches the comprehensive nature of the one developed in this study.

3 Data description and general variables of each policy

A sample dataset has been created from a Spanish insurance company specialized in non-life insurance. The dataset is presented as a spreadsheet file covering the main operations of the company during a period of three completed years (November 2015–December 2018), containing several variables. It includes complete information related to the unique characteristics of different lines of motor insurance, diverse sales and distribution channels, important client-specific details, technical factors, and historical claims data. Each row represents an insurance policy that is

automatically renewed annually until either the policyholder decides to cancel it, or the insurance company terminates it due to non-payment. The database tracks both the addition of new policies and the removal of existing ones from the insured's portfolio. It includes variables to assess the risk, such as the vehicle's price or engine capacity, as well as economic factors impacting the income statement, such as claim costs and insurance premiums.

A comprehensive quality control process has been conducted to ensure the reliability of the dataset. Steps have been taken to address missing, duplicated, or erroneous data, as well as to amend the presence of zero values with respect to the absolute frequency of each variable. This filtering stage holds significant importance for subsequent analyses. The procedure of extracting information from the dataset can be likened to a distillation process, involving a series of sequential stages. Within this process, the initial phase involves tasks such as data cleaning, conditioning, and analysis of the database. These steps ensure the dataset's integrity and its suitability for subsequent analysis. Additionally, the process includes transforming or creating new variables based on existing information. It may also entail adjusting date formats and determining qualitative variables as factors, allowing for further analysis and interpretation.

Despite all the processing, readers interested in the raw/initial database should note that it is available upon request from the authors. We consider this valuable because understanding the quality of a real (non-cleaned) database can serve various purposes. For example, it allows for measuring potential impacts on operational risk in Solvency II due to human errors in data processing or studying the quality of the underwriting process. Furthermore, it can also serve as a practical exercise in data processing and cleaning for students.

Each entry in the dataset represents a distinct insurance contract or policy, with varying maturity dates. Consequently, each policyholder can have multiple rows in the dataset, corresponding to annual observation periods, except for policies initiated less than a year prior to the valuation date. Utilizing an internal identification code (*ID*, see Table 1), all policyholder-product transactions can be accurately recorded. The monetary values considered, such as premiums, are provided after deducting applicable taxes. The dataset comprises 105,555 rows and 30 columns, as indicated in Table 1. Each row in the dataset represents a policy during a period, and each column corresponds to a specific variable.

The variable *ID* corresponds to a unique identification code that is assigned to the combination of policyholder and product. For instance, if there are two rows in the dataset with the same *ID*, it indicates that the policyholder is insured for two distinct time periods (years) while maintaining the same product. The subsequent group of variables, namely *Date_start_contract*, *Date_last_renewal*, and *Date_next_renewal*, are associated with important dates relevant to the insurance policy. Among these variables, *Date_start_contract* represents the start date of the policyholder's contract, *Date_last_renewal* indicates the date of the policyholder's most recent renewal, and *Date_next_renewal* represents the upcoming renewal date for the insurance policy. *Date_birth* is less than 31/12/2000 and greater than 01/01/1918 and the variable *Date_driving_licence* indicates the time when the legal driving license was obtained, with the minimum age requirement being 18 years in

Table 1 First set of variables in the dataset

Variable	Description
<i>ID</i>	Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
<i>Date_start_contract</i>	Start date of the policyholder's contract (DD/MM/YYYY).
<i>Date_last_renewal</i>	Date of last contract renewal (DD/MM/YYYY).
<i>Date_next_renewal</i>	Date of the next contract renewal (DD/MM/YYYY).
<i>Distribution_channel</i>	Classifies the channel through which the policy was contracted. 0 for Agent and 1 for Insurance brokers.
<i>Date_birth</i>	Date of birth of the insured declared in the policy (DD/MM/YYYY).
<i>Date_driving_license</i>	Date of issuance of the insured person's driver's license (DD/MM/YYYY).
<i>Seniority</i>	Total number of years that the insured has been associated with the insurance entity, indicating their level of seniority.
<i>Policies_in_force</i>	Total number of policies held by the insured in the insurance entity during the reference period.
<i>Max_policies</i>	Maximum number of policies that the insured has ever had in force with the insurance entity.
<i>Max_products</i>	Maximum number of products that the insured has simultaneously held at any given point in time.
<i>Lapse</i>	Number of policies that the customer has cancelled or has been cancelled for nonpayment in the current year of maturity, excluding those that have been replaced by another policy.
<i>Date_lapse</i>	Lapse date. Date of contract termination (DD/MM/YYYY).
<i>Payment</i>	Last payment method of the reference policy. 1 represents a half-yearly administrative process and 0 indicates an annual payment method.
<i>Premium</i>	Net premium amount associated with the policy during the current year.
<i>Cost_claims_year</i>	Total cost of claims for the insurance policy during the current year.
<i>N_claims_year</i>	Total number of claims incurred for the insurance policy during the current year.
<i>N_claims_history</i>	Total number of claims filed throughout the entire duration of the insurance policy.
<i>R_Claims_history</i>	Ratio of the number of claims filed for the specific policy to the total duration (whole years) of the policy in force. It provides an indication of the policy's claims frequency history.

Spain. The modal year for this last variable is 2006. The format of all dates in the dataset is DD/MM/YYYY, where DD, MM, and YYYY represent the day of the month, month of the year, and year, respectively.

The *Distribution_channel* classifies the method by which the company got the policy, with a value of 0 indicating an agent channel and a value of 1 representing insurance brokers. The agent channel refers to policies contracted directly through an agent, while the insurance broker channel refers to policies contracted through independent insurance brokers: 60% of the policies are coded with a value of 0 for this variable.

The following group of variables describes the policyholder's affiliation with the insurance company. Figure 1 illustrates the distribution of variables *Seniority*,

Policies_in_force, *Max_policies* and *Max_products*. Specifically, the variable *Seniority* (Fig. 1 left upper panel) indicates the length of time, in years, that the policyholder has been associated with the insurance company. It represents the policyholder’s seniority or tenure within the company with an average duration exceeding 6 years. The variable *Policies_in_force* (Fig. 1 right upper panel) quantifies the number of policies held by each policyholder within the company. On average, each policyholder maintains more than 1 policy with the company. The variable *Max_policies* (Fig. 1 left lower panel) build upon the previous information by measuring the maximum number of products that each policyholder has held within the company. The difference between *Max_policies* and *Policies_in_force* represents the lapse in the portfolio over time. It captures the extent to which the policyholder has reduced the number of products they have with the company, indicating a potential contraction in their insurance coverage or a shift in their needs and preferences. Finally, the company offers a total of 4 different products (car, household, commerce and personal accident) for its policyholders. However, a significant majority of the policyholders, more than 99% of them, have had a maximum of 2 products (see, right lower panel in Fig. 1).

The variables *Lapse* and *Date_lapse* deserve special attention in every insurance portfolio. When a policyholder stops paying premiums and the account value of the insurance policy has already been exhausted, the policy lapses. *Date_lapse* only activates under that event. When this happen, the company has the option to terminate the contractual relationship and the insured can make an independent decision

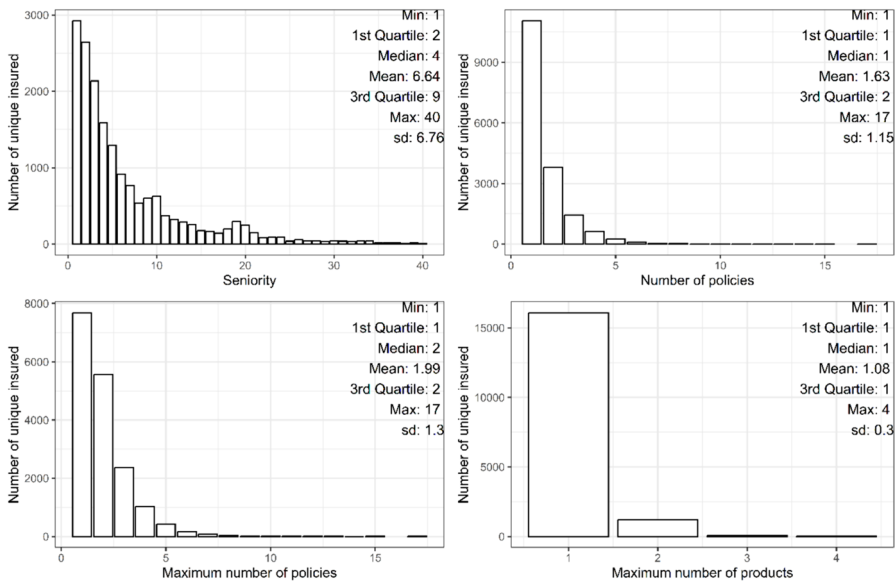


Fig. 1 A graphical summary of policyholders’ affiliation with the insurance company. Top left panel displays the distribution of the number of years of seniority. Top right panel shows the distribution of the number of policies that policyholders have with the insurance company. Lower left panel presents the distribution of the maximum number of policies per insured while the lower right panel exhibits the maximum number of products held by policyholders. In all panels, the year of the last renewal is 2018

of no renewing, citing various reasons for cancelling the coverage. These reasons may include, among others, no longer requiring the product, finding better prices from competitors, or the product not offering the necessary guarantees. *Lapse* indicates either non-permanence or permanence and can be used to measure the number and the date of policies that the company retains (does not retain). Somewhat related to these variables is the *Payment* variable, which captures the frequency of the payment method. A value of 0 indicates an annual payment method and a value of 1 a semi-annual payment method.

The following subset of variables collects quantitative variables that are related to income and expenses linked to the portfolio. The *Premium* variable (income) quantifies the amount received by the insurer in connection to a policy during the current year after deducting taxes but including both internal (administrative) and external (commercial) expenses. Figure 2, left panel, shows a plot density of the *Premium* variable, which presents an asymmetric (positive) distribution. On the other hand, Fig. 2, right panel, illustrates the distribution of *Cost_claims_year* variable, which exhibits a comparable pattern to premium figures after excluding, for presentation purposes, policies with 0 claims. The *Cost_claims_year* variable includes both payments and reserves, with the total claim cost being the sum of these two components. The main statistics are accompanied by the right-hand panel, which also displays the average cost per claim (mean claim). This average is calculated by adding all amounts together and dividing them by the total number of claims. All the figures are in euros.

Finally, the last subset of variables in Table 1 describes characteristics related to the policy and the number of claims. These variables may encompass various factors

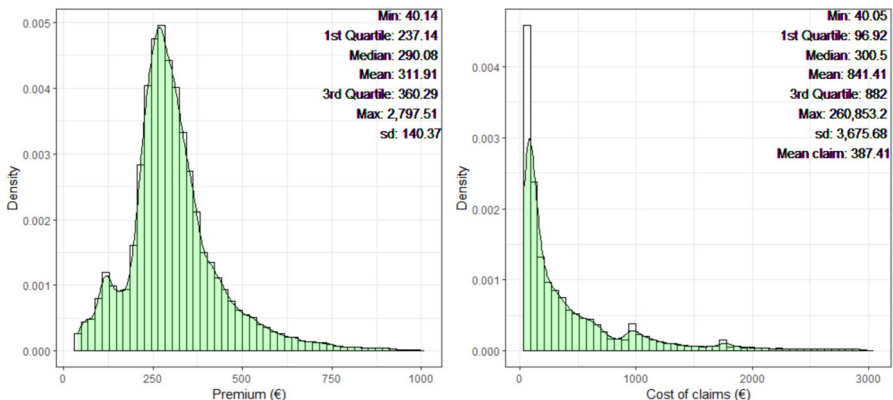


Fig. 2 Graphical representation of income and expenses distributions. A density estimation of the amount of insurance premiums is displayed for the range 0–1000 in the left panel. In the right panel, a density estimation of the cost of policies that have experienced a claim is presented for the range 0–3000, but excluding claims linked to CICOS. The CICOS agreement in Spain facilitates handling claims from vehicle collisions, aiming to expedite compensation for material damage. It sets a predetermined amount agreed upon by the involved insurance companies, regardless of the actual claim, with costs rising from around €800 in 2019 to over €1000 in 2023. In both panels, the year of the last renewal is 2018. As can be observed, the summary statistics included in each panel have been calculated without imposing any restrictions on the upper amounts

and attributes that can influence or provide insights into the claim's behaviour of the insured. The variables N_claims_year and $N_claims_history$ represent the number of claims that the insured has made during the current year and throughout the entire history of the policy, respectively. The presence of zero values in any of the above variables implies that the policy has not incurred any claims. The $R_Claims_history$ variable is defined as the ratio of the number of claims filed for a particular policy to the total duration of the policy in force, measured in whole years. This ratio serves as an indicator of the policy's historical claims frequency. Note that the database was extracted in 2019 and includes claims up to 2018. Therefore, the most recent claims exhibit a run-off with pending developments that will change slightly over time.

4 Deeping on the granularity of the information: disaggregating by type of claim

Despite the high level of detail in the described database, some variables, such as the number and cost of claims, are presented in an aggregated form for all types of claims. As stated by one of the reviewers, this aggregation hides some relevant information, and separating the claims by type would add much value. Therefore, in order to address this issue, in this section, we describe a second dataset that presents more granular information for a sample of policies. Unfortunately, we were not provided with all the details for the entire set of policies. The company considers it risky to reveal the full set of information for all the policies. According to them, this could expose a significant part of its underlying distributions of risk, due to the specific nature of this information, which determines the full granularity of the risk. Hence, the company has provided us with only a sample of the types of claims in the database. This subset includes detailed information on the nature of each claim, along with a more specific breakdown of the amount for each claim. Although it is just a sample, representing around 15% of claims, this information is of great interest for determining the composition of some of the most significant risk variables.

This information is available in the secondary file *sample_type_claim.csv*, also at the URL: <https://doi.org/10.17632/5cxyb5fp4f.2>. The file consists of four columns in total. The first two columns, *ID* and *Cost_claims_year*, represent the identifier for each insured and the total cost in the generic database, respectively. The third column, *Cost_claims_type*, specifies the amount of the claim by type. In this case, the sum of the *Cost_claims_type* amounts for each *ID* corresponds to the *Cost_claims_year* in the primary database. Note that *Cost_claims_year* is included only for identification purposes and that its value is repeated for an *ID* with multiple claims, so the total sample sum does not hold any economic significance.

The final column corresponds to the variable *Claims_type*, which informs about the type of claim linked to the amount detailed in *Cost_claims_type*. We have identified up to nine types of claims in the subset. *Travel assistance* and *broken windows* are the claims with the smallest amounts of money, causing the least damage to the vehicle. The category *complaint* means that the insured driver is innocent and is seeking compensation for damages from a third party. The option *negligence* means that the insured

driver is at fault in the accident, such as by hitting another vehicle. There are also classic claims such as *theft* and *fire*, with fire having one of the highest average amounts in the sample. Another category includes *all risks* that means insured person's own damages are compensated, minus any applicable deductibles. The *injuries* category refers to damage caused to occupants and represents the type of claim with the highest average amount. Lastly, there is a diverse category, although much less frequent, labelled as *other*, which includes various types of claims, such as those caused by hitting animals.

5 Factors determining risk associated of vehicles/insured

The previous variables (Table 1) establish the primary characteristics of the policy and serve as the main economic indicators. Table 2 set of variables, available in the dataset just after the ones described in Table 1, enables a comprehensive analysis of the risk undertaken by the insurance company.

The variable *Type_risk* allows the company to distinguish between various insured risk profiles: 1 represents motorbikes (8502 rows, 8.1%), 2 corresponds to vans (13,212 rows, 12.5%), 3 denotes passenger cars (82,990 rows, 78.6%), and 4 pertains to agricultural vehicles (851 rows, 0.8%). The risk profiles are insured based on their predominant *Area* of circulation, with the rural area accounting for over 70% of the policies. Additionally, there is the possibility of including multiple regular drivers (*Second_driver*) in the policy. Specifically, over 10% of the policies include at least a second regular driver.

The last set of variables of Table 2 defines the specific and technical characteristics of the insured vehicles. Firstly, *Year_matriculation* shows a prevalence of the years 2004–2007, reflecting the economic expansion before the onset of the financial crisis in 2008. The variables *Power* and *Value_vehicle* exhibit a wide range of numerical values and may demonstrate a degree of linear correlation, depending on the specific variable under analysis. Figure 3 illustrates the relationship between these two variables based on the *Type_risk* variable. *Power* is also typically correlated with *Cylinder_capacity* due to the power of a vehicle's engine often corresponds to its cylinder capacity, indicating a strong relationship between these two factors.

Continuing the risk description, five new variables provide technical details about the vehicles. For motorbikes and agricultural vehicles, the standard value for *N_doors* is 0, whereas for vans and passenger cars, it is typically 5. Within the database, two different fuels are identified with the variable *Type_fuel*: petrol, accounting for 36.8% of the rows, and diesel, comprising 61.6%. Additionally, there is a total of 1.7% categorized as "NA" indicating that the fuel type is not available for these entries. The variables *Length* and *Weight* enable the determination of the risk size. Averages of *Length* and *Weight* are the 4.25 m and 1,191 kg, respectively.

6 Unlimited future case studies

The database presented in this paper offers the opportunity to open a large number of possible studies. One block of studies may focus on predicting variables based on the behaviours of others, such as those employing machine learning techniques.

Table 2 Second set of variables in the dataset

Variable	Description
<i>Type_risk</i>	Type of risk associated with the policy. Each value corresponds to a specific risk type: 1 for motorbikes, 2 for vans, 3 for passenger cars and 4 for agricultural vehicles
<i>Area</i>	Dichotomous variable indicates the area. 0 for rural and 1 for urban (more than 30,000 inhabitants) in terms of traffic conditions.
<i>Second_driver</i>	1 if there are multiple regular drivers declared, or 0 if only one driver is declared.
<i>Year_matriculation</i>	Year of registration of the vehicle (YYYY).
<i>Power</i>	Vehicle power measured in horsepower.
<i>Cylinder_capacity</i>	Cylinder capacity of the vehicle.
<i>Value_vehicle</i>	Market value of the vehicle on 31/12/2019.
<i>N_doors</i>	Number of vehicle doors.
<i>Type_fuel</i>	Specific kind of energy source used to power a vehicle. Petrol (P) or Diesel (D).
<i>Length</i>	Length, in meters, of the vehicle.
<i>Weight</i>	Weight, in kilograms, of the vehicle.

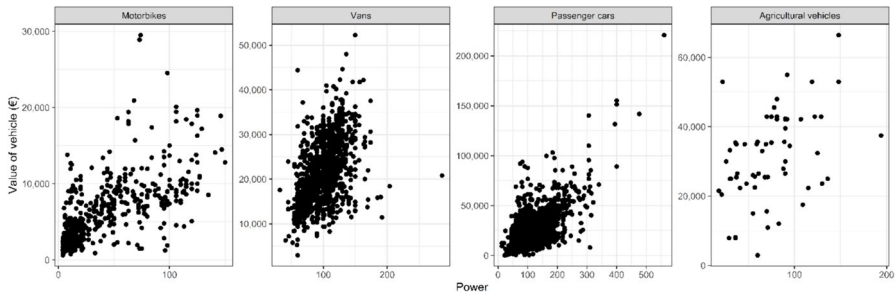


Fig. 3 Relation between *Power* and *Value_vehicle* categorized by *Type_risk*

Variables like claims or vehicle values can be explored in relation to other factors. Within this category, classical linear regression or logistic regression models can be used, as well as decision trees, support vector machines or artificial neural networks. The dataset could be used by actuarial students for exploring and comparing different strategies of pricing and can be used by trainers for creating, by conveniently splitting the dataset, proper challenges for testing the students' ability to rule a motor insurance business.

Another critical aspect of top importance of risk management lies in its impact on the various sub-modules within the comprehensive Solvency II modular framework. Among these, the underwriting policy emerges as a fundamental pillar, having profound influence in risk selection strategies. Insurers are tasked with strategically aligning their approaches to increase profitability by grouping their key policyholder segments or those contributing the highest potential value [6]. Consequently, segmentation facilitates the design of strategies aimed at retaining policyholders with minimal claims history while implementing premium adjustments for those with a higher claim's frequency. The present database allows clustering algorithms to segment the portfolio into different risk profiles.

Aligned with the modular risks of Solvency II, operational risk—particularly the aspect associated with human error—stands out as one of the least explored risks in actuarial literature, likely due to the challenge of accessing real data for analysis. Upon request to the authors, readers can access the raw database to examine the typology or structure of various errors within the data, such as missing values or out-of-range values. Furthermore, the repercussions of these errors on critical factors like premiums or claims can be thoroughly analysed.

Lapse or Lapse rate have emerged as one of the most studied key performance indicators (KPIs) within insurance companies in recent years, particularly due to the current challenges of increasing inflation, which delay the ability to raise premiums. The database conveniently presents two columns: one for portfolio decline and the associated date. We encourage researchers to delve into various aspects, such as identifying the variables influencing portfolio decline or indicative the specific timing when it materializes.

Lastly, and not less important, the database facilitates the execution of classical exploratory data analysis (EDA) studies, offering a valuable opportunity to develop

essential skills in initial data processing and variable analysis. Particularly beneficial for numerous university studies, this database enables the exploration of numerical, categorical and date variables, providing a deeper understanding of their distributions and extracting various basic statistics. Moreover, it serves as an ideal platform for understanding both linear and non-linear relationships between numerical and categorical variables, developing a comprehensive learning experience.

7 Conclusions

The dataset presented in this paper comprises information from an insurance company in Spain, spanning the period from 2015 to 2018. The insurance policy is renewable annually and is automatically renewed until either the insured opts to cancel or the company cancels due to non-payment. The database inherently includes records of new business entries and cancellations within the insured's portfolio. The reference policy's frequency and claim costs consist of various claim types, such as material damage, body injury, vehicle all-risk coverage, travel assistance-related claims, glass amounts, and vehicle theft. Regarding the insured person's premiums, these are comprised of the total sum of payments for different guarantees and products included in the policy.

Apart from the variables available in the dataset, additional variables necessary for various analyses, such as client profitability, can be derived. Client profitability can be determined by considering the premiums paid during a specific period and the claims incurred, considering both closed file (claims already settled) and provisioned file (claims yet to be settled). Non-life insurance often involves calculating the claims frequency, which is the ratio of the number of claims to the exposure period of the insured contract in days. Another significant calculation is the claims ratio, expressed as a percentage, which indicates the portion of premium consumed by the claims experience. A ratio exceeding 100% signifies losses in the technical aspect of the business.

Each of the mentioned economic indicators can be applied to analyse and measure the variables that influence claim behaviour. Moreover, these indicators can be utilized to classify and group variables based on their detection patterns using machine learning techniques. By employing classification methods, we can categorize the variables effectively, allowing for a deeper understanding of their impact on the claims process. These techniques offer valuable insights into the complex relationships between the variables and the overall insurance claims behaviour. Finally, it is worth noting that while this database provides valuable insights, its coverage might not be exhaustive and may have certain limitations. As the insurance industry increasingly embraces data-driven approaches and digital transformations, the expansion and enrichment of databases, along with the inclusion of other lines of business, become essential for insurers to make more informed decisions, create personalized policies, and enhance risk management strategies.

Acknowledgements The authors wish to thank the editor and two reviewers for really valuable suggestions and comments. The authors acknowledge the support of Ministerio de Ciencia e Innovación through

project PID2021-128228NB-I00 and the support of Generalitat Valenciana through project CIGE/2023/7 (Conselleria de Educaci3n, Universidades y Empleo).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare to have no known competing financial interests or personal relationships that may have influenced the work reported in this paper.

Ethics statements The relevant informed consent was obtained by the company from the insured in the moment of contracting the product. Data is offered anonymised.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Department of Health and Human Services (2021) Health insurance marketplace: summary enrollment data for the initial annual open enrollment period. ASPE. <https://links.uv.es/TLbW7X5>
2. Lled3 J, Pavia JM (2022) Dataset of an actual risk-life insurance portfolio. Data Brief 45:108655
3. New York State Department of Financial Services (2024c) Total property insurance premiums written annually in New York: Beginning 1998. NY Open Data. <https://links.uv.es/bhc9akR>
4. New York State Department of Financial Services (2024b) Motor vehicle point & insurance reduction program (PIRP) participation: Five-year window. NY Open Data, 2022. <https://links.uv.es/qYWU9Tg>
5. New York State Department of Financial Services (2024a) DFS automobile insurance company complaint rankings: beginning. NY Open Data, 2022. <https://links.uv.es/IQtNU63>
6. Payne A, Frow P (2005) A strategic framework for customer relationship management. J Mark 69(4):167–176
7. Venkatesan R, Kumar VA (2004) Customer lifetime value framework for customer selection and resource allocation strategy. J Mark 68(4):106–125

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.