Journal of Corporate Governance, Insurance, and Risk Management
https://www.acadlore.com/journals/JCGIRM

Modeling Vehicle Accident Risks in Auto Insurance: An Application of Generalized Linear Models in the Context of the National Insurance Company, Regional Directorate of Setif, Algeria

Chahrazed Salhi, Djamal Tebache*

Department of Finance and Accounting, Ferhat Abbas University Setif-1-, 19000 Setif, Algeria

* Correspondence: Djamal Tebache (djamal.tebache@univ-setif.dz)

Received: 08-03-2024 **Revised:** 09-15-2024 **Accepted:** 09-24-2024

Citation: Salhi, C. & Tebache, D. (2024). Modeling vehicle accident risks in auto insurance: An application of generalized linear models in the context of the National Insurance Company, regional directorate of Setif, Algeria. *J. Corp. Gov. Insur. Risk Manag.*, *11*(3), 180-198. https://doi.org/10.56578/jcgirm110304.

Abstract: This study investigates risk distribution models in the context of auto insurance in emerging markets, with a focus on the National Insurance Company (SAA), regional directorate of Setif, Algeria. The research applies generalized linear models (GLM) and factor analysis to model the frequency of vehicle accidents and their associated risks. A comprehensive approach is employed, beginning with a discussion of the techniques used for data collection and preliminary descriptive analysis. Following this, a theoretical framework is established for understanding the risk distribution models, highlighting the role of GLM in the modelling of accident frequencies within the insurance industry. Different types of factor analysis, including basic coefficient analysis, cross-factor analysis, generalized cross-factor analysis, and mixed factor analysis, are examined in relation to their applicability to insurance risk modelling. Subsequently, generalized linear models are implemented to derive a robust model for accident frequency, utilizing R software for analysis. The results reveal that the pricing system of the National Insurance Company is influenced by multiple, non-deterministic factors, which complicate the prediction of accident rates and insurance costs. These findings underscore the importance of incorporating various risk factors into pricing strategies, rather than relying on deterministic models. The study highlights the necessity of considering a broader range of factors in the development of pricing systems, particularly in emerging markets where data may be incomplete or subject to considerable variability. Furthermore, the use of Mixed Poisson models is suggested as an effective approach for capturing the non-linear relationship between various risk factors and accident occurrence. This research contributes to the existing body of knowledge by providing a nuanced understanding of the application of GLM and factor analysis in the auto insurance sector, particularly in emerging markets.

Keywords: Auto insurance; Pricing; Generalized linear model (GLM); Factor analysis; Mixed Poisson model

1. Introduction

In recent years, Algeria has seen a sharp rise in the frequency of car accidents: more than 32,200 accidents in 2022, more than 636,697 in 2023, and 12,162 accidents in the first half of 2024, according to the National Road Safety Delegation. Tens of thousands of traffic accidents are reported to the appropriate authorities (National Gendarmerie, Civil Protection, etc.) each year, requiring a significant investment of time and money. Consequently, numerous files are presented to insurance companies seeking reimbursement. Due to the delay in paying out compensation, this has a detrimental effect on both the financial performance of these businesses and consumer loyalty.

The insurance business uses the theory of probability to compute losses, which is the basis for determining premiums, because it is uncertain to estimate the incidence of traffic accidents due to randomness. The distribution of accident rates must be understood in order to develop a deterministic method for pricing auto accidents. Since the number of accidents is random, we need to know how accidents are distributed. Any combination of variables that follows the Bernoulli distribution leads to the binomial distribution; however, if the number of observations is large and the probability of success is small, the binomial distribution will approximate the Poisson distribution

as a mathematical expectation. For a single driver, it is a random variable that takes the values 1 if the accident occurs and 0 if it does not; we have Bernoulli's law with probability (Denuit et al., 2007). However, a single driver can cause multiple accidents, and the insurance company's portfolio contains a large number of insurance policies.

Hence, the Poisson distribution occupies the main role in modeling independent and counting data because it is adapted as a model where there is only randomness and within a homogeneous population. However, these two conditions are not always met in the case of modeling insurance-related data (Veysseyre, 2007). Therefore, we resort to Mixed Poisson models (negative binomial) to describe populations consisting of an infinite number of homogeneous subpopulations. In this research, we are trying to answer the following question: How is the frequency of vehicle accidents distributed in Algeria?

This leads us to test the following hypothesis: The frequency of vehicle accidents in Algeria follows a negative binomial law.

For this end, we began with a literature review, and then we should focus on a theoretical examination of risk distribution models. Additionally, we used generalized linear models and factor analysis to model the number of accidents. Before delving into the specifics of the generalized linear models for the study data, we provided a theoretical analysis of the last two concepts by outlining the various types of factor analysis, including basic coefficient analysis, cross factor analysis, generalized cross factor analysis, and mixed factor analysis.

2. Literature Review

As Lemaire (1985) remarked, of all types of non-life insurance, automobile third party undoubtedly gives rise to the most heated debate. This type of study has been a source of interest to many scholars, who have investigated some of the elements mentioned in this study, including:

In the first study, Ghali (2002) examined a marginal pricing model for auto insurance in a regulated market. The study's aim was to use the marginal model to analyze the Tunisian auto insurance pricing system; to do this, the researcher employed a pricing model based on before and after characteristics. The study was carried out at the level of a significant private Tunisian company that held 7% of the country's auto insurance market between 1990 and 1995. The study's sample consisted of 46,337 observations that were distributed annually during the same time period. From this sample and using counting models (Poisson and negative binomial), the importance of factors explaining vehicle accidents was estimated from the annual data, as well as the formation of marginal bonus and penalty tables (bonus-malus).

The findings were that the Tunisian pricing system and the reward and penalty system are not marginal, as evidenced by the presence of other variables in addition to puissance and usage that are significant and explain the number of accidents.

A study on pricing and segmentation in auto insurance was carried out on the French insurance company Mutant d'Assurance by Guillaume (2010). The study examined the company's operations in 2008 by examining all auto insurance policies with at least one day of guarantee during that year, as well as the losses that were reported during that same year. The following models were used in this study to develop models for the number of accidents and the amount of losses independently using generalized linear models (GLMs) after the data had been processed and corrected using factor analysis: Poisson, quasi-Poisson and gamma. The results obtained through this study are as follows: Building a pricing model for auto insurance based on generalized linear models and presenting a methodology for segmentation analysis in auto insurance pricing.

The third study examined the factors that influence the frequency of losses in auto insurance and was conducted by Vasechko et al. (2009). This study was conducted on a sample of 50,000 observations in a French insurance company; the data of this sample are 4-wheel tourist cars during the year 2005. Typically, counting models (Poisson and negative binomial models) are used to model accident frequency in this study, which aims to identify the factors that explain the number of liability accidents reported by the insured to the insurer. However, a significant portion of the insureds in the insurance portfolio may have no recorded losses during the insurance period (year); this zero value may indicate either no loss or no declaration. In order to capture the importance of these null values as well as the heterogeneity in the study population, zero-inflated Poisson (ZIP) and zero-inflated Poisson (ZIP) models that follow a non-binomial distribution (ZINB) were used. Variables explaining the frequency of losses are the same as in classical counting models, except that the choice of contract suggests an adverse selection effect. Findings are organized as follows: Results related to the Poisson model and the negative binomial model; results related to the ZIP and ZINB models; and then the comparison between the models.

Regarding the Poisson model and the negative binomial model, the study demonstrated that both models generated the same significant variables with comparable outcomes. The relationship between these variables was as follows: the type of driver, i.e., whether the insured is the same driver or not; the damage guarantee on the car in the three cases (with an important exemption, medium exemption, and weak exemption); the age of the car; the age of the license; and the bonus and penalty factor. While the other elements reduce the likelihood of recurring losses, the age of the vehicle and the reward and penalty factors increase it. However, because of the over dispersion in the data, the negative binomial model is a better fit for the data than the Poisson model.

Since it has already been stated how important it is for insureds to have no allowed losses (i.e., zero documented losses), the aforementioned models (ZIP and ZINB) were used, and the following outcomes were obtained:

The first part, which deals with the frequency of losses, yields the same results as the classical counting models. The second part, which deals with data that has been inflated with zero values, shows that the probability of loss decreases according to the bonus and penalty coefficient and increases with the age of the car, the driver's license, and the damage guarantee on the car with an average exemption. To compare all these models, the Vuong test was used, and the result obtained was that a ZIP model is favored over the standard Poisson model, and a ZINB model is favored over a negative binomial model. Finally, ZINB is recommended as the final model.

In the following study, Lai (2011) examined the development of a model to assess the risk of traffic accidents in urban areas using the Structural Equation Model. Because there are many variables influencing the occurrence of accidents, the researcher restricted them to three dimensions: driver characteristics, vehicle characteristics, and road characteristics. As a result, the study variables were determined to be as follows: The dependent variable: The risk of road traffic accidents, including both the risk realization rate, i.e., the number of accidents in relation to the number of vehicles, and the severity of the risk in relation to the number of deaths and injuries.

The explanatory variables, in turn, include driver characteristics (gender, age, license, and blood alcohol content); vehicle characteristics (vehicle type and traffic volume); and road characteristics (road width, road straightness, etc.).

The applied study was conducted in Taiwan Province (People's Republic of China) with a coverage of 26 roads divided into 249 segments according to road characteristics; the data on accidents were obtained from the police office, and the study period lasted from 1 January to 31 December 2003.

He drew the following conclusions: Both the driver and road dimensions have the greatest impact on the realization of risk, while the car dimension is not statistically significant, and concluded that the most influential cause of accidents always remains human.

Locating the current study within previous studies: It is clear from the presentation of earlier research that a Poisson model and a negative binomial are required to model the number of vehicle accidents. Nevertheless, the data shows that there are many insured individuals with zero accidents, which may be due to the insured failing to report the accident or to other factors. To account for this situation, the author employed ZIP and ZINB models. Therefore, this study differs from previous studies in several points that we summarize:

In terms of work environment: Previous studies were conducted in foreign countries such as France, Taiwan, and Arab countries (Tunisia). The current study is concerned with the case of Algeria, considered as an emerging market.

In terms of objectives, the current study uses factor analysis to process and rectify the data in order to estimate the number of car accidents that occurred in the 2023/2024 period. It then uses generalized linear models to develop ZIP and ZINB models for the number of losses. The models listed above.

2.1 Mixed Poisson Regression Models

The end of the 1970s saw the introduction of this kind of model into modern economic analysis, where the dependent variable is valid and non-negative in order to account for its quirks. Numerous studies have been conducted in this area, such as those by Hausman et al. (1984) on the number of races in a specific time period; Cameron & Trivedi (1986) on the number of doctor examinations; Dionne & Vanasse (1989) on the number of accidents; and Winkelmann (1995) on the frequency of job changes.

In general, a Poisson regression model can be defined as a model that relates a discrete dependent variable, which takes positive and valid values, to one or more explanatory variables. In order to use this model, the following conditions must be met:

- The instantaneous probability of loss is proportional to the length of the period under consideration.
- The instantaneous probability of an event occurring is constant over the period under consideration (i.e., the risk is constant in time).
- The probability of more than one event occurring is low.
- Accidents are independent of each other.

These conditions are consistent with the Poisson distribution, which is characterized as the law of instantaneous and independent accidents; if the probability of an individual being involved in an accident during a given period is equal to (Partrat & Besson, 2005):

$$P(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \tag{1}$$

The Poisson parameter we are trying to estimate, which expresses both the mean and variance of the distribution. If the Poisson distribution follows any $N \rightsquigarrow P(\lambda)$, then the properties of the Poisson distribution can be summarized

as follows (as a Moment generating function):

$$\varphi_N(z) = \sum_{k=0}^{\infty} e^{zk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(e^z \lambda\right)^k}{k!} = e^{-\lambda} e^{\left(e^z \lambda\right)} = e^{\lambda\left(e^z - 1\right)}$$

Moments: From $\mu_{[k]} = \lambda^k$ we conclude that:

$$m = E(N) = \lambda$$

$$\sigma^2 = V(N) = E\left(N^2\right) - \left[E(N)\right]^2 = \left(\lambda + \lambda^2\right) - \lambda^2 = \lambda$$

$$\gamma_1 = \frac{1}{\sqrt{\lambda}},\ \gamma_2 = \frac{1}{\lambda}$$

The equality between the mean and variance, which we express as equi-dispersion, is the most important characteristic of the Poisson distribution.

Additivity: Given two independent random variables N_1 and N_2, where: $N_1 \rightsquigarrow P(\lambda_1)$ and $N_2 \rightsquigarrow P(\lambda_2)$ then:

$$N_1 + N_2 \rightsquigarrow \mathrm{P}(\lambda_1 + \lambda_2)$$

Approximation towards the normal law $N(\lambda, \lambda)$ with continuity correction:

$$P(N \le n) \simeq \Phi\left[\frac{n + 1/2 - \lambda}{\sqrt{\lambda}}\right]$$

Criteria Used: $\lambda > 18$

Decomposition: If the loss is divided into r independent categories (exhaustive), e.g., r is warranty, then if the total frequency N of losses follows Poisson's law $P(\lambda)$, and if the frequency for each i (where $i = 1 \ldots, r$). We have a probability that the loss is from the category of i, and Ni is the frequency of loss corresponding to that category. Random variables are independent variables distributed according to Poisson's laws $P(\lambda\pi_1)$, $P(\lambda\pi_2)$, ..., $P(\lambda\pi_r)$.

Statistical inference: If (n_1, n_1, ……, n_1) represents the sample observations (N) of the random variable, both the moments method and the maximum likelihood method lead to the same estimator for λ, which is $\hat{\lambda} = \bar{n}$, t is evident from the foregoing that the conventional linear regression model is ineffective for two reasons when representing the relationship between a discrete dependent variable and explanatory variables: First of all, the shape of the cloud of observations does not fit the linear form. Second, the Poisson distribution is consistent with these two assumptions, suggesting that the normality hypothesis may not be negative. However, because of the high number of missing values and the existence of some extreme values, the hypothesis of equality between mathematical expectation and variance, which implies the homogeneity of the portfolio with regard to risk, is severely constrained (de Jong & Heller, 2008). Here, the variance exceeds the mean; we discuss a concept known as overdispersion of the variable N. This scenario necessitates a partial standard deviation estimation, which could lead to the null hypothesis regarding the estimator vector's dependability β in the model being rejected. By adding a new parameter that might represent the unobserved heterogeneity of the implicit variables that may contribute to this over-dispersion, the objective is to employ a different model that accounts for this over-dispersion. Mixed Poisson models are the models that address this problem.

2.1.1 Mixed Poisson models

Since each driver or group of drivers has unique characteristics that influence the likelihood of traffic accidents, the society we are studying (the population of insured drivers) is heterogeneous. For this reason, we use Mixed Poisson models to examine the distribution of losses (Lee et al., 2018).

Definition of the Mixed Poisson model: The Poisson distribution often shows an inadequate fit to the observations of the underwriter's portfolio due to heterogeneity among underwriters; here we multiply the average loss frequency λ by a positive random variable Θ, so the average loss realization becomes a random variable dependent on Θ. We choose Θ, so $E(\Theta)=1$ because we want to approximate the same frequency of occurrences in the portfolio under our condition Θ, we have:

$$\Pr\left[N = k / \Theta = \theta\right] = p\left(k / \lambda\theta\right) = \frac{e^{-\lambda\theta}\left(\lambda\theta\right)^k}{k!},\ k = 0,1,... \tag{2}$$

where, $p(\lambda\theta)$ is the Poisson probability function with mean $\lambda\theta$. This approach means that not all underwriters necessarily have the same accident frequency, but some have a high average ($\lambda\theta$ where $\theta \geq 1$) and some have a low average ($\lambda\theta$ where $\theta \leq 1$).

The probability of vehicle accidents, as mentioned earlier, mostly follows Poisson's mixed law, so the probability of registering a K number of losses with the insurance company as a result of these accidents has the same formula (2) with Θ expectation. In the general case, the Θ variable is random, neither continuous nor discrete, but a combination of both, so the probability distribution is as follows:

$$\Pr\left[N = k\right] = E\left[\ p\left(k / \lambda\Theta\right)\right] = \int_0^\infty \frac{e^{-\lambda\theta}\left(\lambda\theta\right)^k}{k!} dF_\Theta\left(\theta\right) / F_\Theta \tag{3}$$

wherein F_Θ the distribution function of the variable Θ.

Hence, we say that the random variable N follows a Poisson distribution mixed with a parameter λ and level of risk equal to Θ, with a probability function of the formula (3), which we denote by $N \rightsquigarrow MPoi(\lambda, \Theta)$.

Become in this case:

$$\varphi_N\left(z\right) = \int_0^\infty e^{-\lambda\theta} \sum_{k=0}^{+\infty} \frac{\left(z\lambda\theta\right)^k}{k!} dF_\Theta\left(\theta\right) = M_\Theta\left(\lambda\left(z-1\right)\right)$$

Rewarding $M_\Theta(t) = \varphi_N\left(1 + \frac{t}{\lambda}\right)$, so we obtain mathematical expectation for Mixed Poisson distribution as follows:

$$E\left(N\right) = \int_0^{+\infty} \sum_{k=0}^{+\infty} k \frac{e^{-\lambda\theta}\left(\lambda\theta\right)^k}{k!} dF_\Theta\left(\theta\right) = \lambda E\left(\Theta\right) = \lambda$$

And a variance

$$V\left(N\right) = \lambda + \lambda^2 V\left(\Theta\right)$$

Note that the variance in this case is larger than in the Poisson distribution, which is justified by the overdispersion:

$$\gamma\left(N\right) = \frac{1}{\left(V\left(N\right)\right)^{3/2}} \left(3V\left(N\right) - 2E\left(N\right) + \frac{\gamma\left(\Theta\right)}{\sqrt{V\left(\Theta\right)}} \frac{\left(V\left(N\right) - E\left(N\right)\right)^2}{E\left(N\right)} \right)$$

2.1.2 Poisson Inverse Gaussian distribution

Here we complete formulas (2) and (3) as $\Theta \rightsquigarrow$ Igau(1, τ) becomes

$$f_\Theta\left(\theta\right) = \frac{1}{\sqrt{2\pi\tau\theta^3}} e^{-\frac{1}{2\tau\theta}\left(\theta-1\right)^2}, \quad \theta > 0$$

And a probability function

$$\Pr\left[N = k\right] = \int_0^\infty e^{-\lambda d\theta} \frac{\left(\lambda d\theta\right)^k}{k!} \frac{1}{\sqrt{2\pi\tau\theta^3}} e^{-\frac{1}{2\tau\theta}\left(\theta-1\right)^2} d\theta \tag{4}$$

Poisson Inverse Gaussian distribution characteristics:
Mathematical expectation

$$E(N)=\lambda$$

Variance

$$V(N)=\lambda+\lambda^2\tau,\ \gamma(\Theta)\Big/\sqrt{V(N)}=3$$

The generating function for moments of *N* is given by the relation:

$$\varphi_N(z)=e^{\frac{1}{\tau}\left(1-\sqrt{1-2\tau\lambda(z-1)}\right)}$$

2.1.3 Poisson log normal distribution
When $\backsim\log N(-\sigma^2/2,\sigma^2)$, taking $\mu=-\sigma^2/2$ (as $E(\Theta)=1$), the probability density function of Θ is

$$f_\Theta(\theta)=\frac{1}{\theta\sigma\sqrt{2\pi}}e^{\left(-\frac{\left(\ln\theta+\sigma^2/2\right)^2}{2\sigma^2}\right)},\quad \theta>0$$

However, the probability function for the Poisson log normal mixed distribution is given by the following formula:

$$\Pr[N=k]=\frac{1}{\sigma\sqrt{2\pi}}\frac{(\lambda d)^k}{k!}\int_0^\infty e^{(-\lambda d\theta)}\theta^{k-1}e^{\left(-\frac{\left(\ln\theta+\sigma^2/2\right)^2}{2\sigma^2}\right)}d\theta \qquad (5)$$

Using relations (5) and (6), we can deduce the mathematical expectation and variance as follows:

$$E(N)=\lambda$$

$$V(N)=\lambda+\lambda^2\left(e^{(\sigma^2)}-1\right)$$

2.1.4 The binomial negative model
The Poisson distribution has been used to express the distribution of accidents for a set of individuals that λ is implicitly assumed to contain all the information to express the probability of an accident occurring, but this property is too restrictive to study this type of model. We first consider the case where λ do not have all the information about the individuals.

A. Count models without individual features
Assuming that λ does not contain all the information and for each individual the number of accidents follows a Poisson law, it is convenient to assume that λ follows Γ distribution of two parameters *a* and τ, in the case of the distribution of the number of losses in auto insurance.

So, λ distribution is *f(λ)*, such as $f(\lambda)=\frac{\tau^a e^{-\tau\lambda}\lambda^{a-1}}{\Gamma(a)}$, with a mean = $\frac{a}{\tau}$ and a variance = $\frac{a}{\tau^2}$. $\Gamma(a)$ is a Gamma function of *a*.
The probability of a randomly selected individual achieving an accident is defined as:

$$p(k/a,\tau)=\frac{\int_0^\infty e^{-\lambda}\lambda^k f(\lambda)d\lambda}{k!}=\frac{\Gamma(a+k)\tau^a}{\Gamma(k+1)\Gamma(a)(1+\tau)^{k+a}} \qquad (6)$$

With an average = $\frac{a}{\tau}$, and a variance = $\frac{a}{\tau}\left(1+\frac{1}{\tau}\right)$.

Therefore, we say that the negative binomial law follows with two parameters $\left(a,\frac{a}{\tau}\right)$, and we write:

$$\text{X}\rightsquigarrow\text{BinN}\left(a,\frac{a}{\tau}\right)$$

The most important use of this type of distribution in general insurance is in the distribution of the number of losses where the risks are heterogeneous and the variance is greater than the mathematical expectation.

B. Count models with individual characteristics and their application

We assume that a variable N_i represents the number of incidents for a person i that occur during a given period. If N_i is independent of N_i for all $\neq i$, then the set of such variables follows a parameterized Poisson's law λ_i.

In the counting models with individual characteristics, the practical formula that relates the parameter λ_i to the individual variables is as follows: $\lambda_i=exp(X_i\beta)$, where β is the vector of the parameters we are estimating that λ_i represents the mean and variance.

Using the exponent allows us to obtain a non-negative mean and variance that cancels out the linear regression models. Also, the probability of an individual i having K_i accident during a given period is given as follows:

$$\Pr\left[N_i=k_i\right]=\frac{e^{-\exp(X_i\beta)}\exp\left(X_i\beta\right)^{k_i}}{k_i!}$$

The maximum plausibility function is given by

$$L\left(N_i,\beta\right)=\prod_{i=1}^{n}\frac{e^{-\exp(X_i\beta)}\exp\left(X_i\beta\right)^{k_i}}{k_i!}$$

Since the logarithmic function is monotonic, this allows for a simple maximization of the logarithm of the maximum likelihood function rather than the function itself, and since the logarithm of the maximum likelihood function is not linear in β, solving the sentence requires the use of an iterative algorithm such as Newton Raphson methods.

$$\beta^{t+1}=\beta^{t}-\left[H\left(\beta^{t}\right)\right]^{-1}g\left(\beta^{t}\right)$$

where, $g(.)$ represents that *gradient* is the logarithm of the plausibility function, β^t is the arbitrary initial value, and iteration process terminates when the convergence condition is met (LIMDEP allows us to easily calculate a β value).

However, the previous formulation suffers from at least two drawbacks: the model is built on the assumption of the independence of successive events with the assumption that the mean and variance of N_i are equal, and the second drawback is that the variables X_i express all the probabilities of the events. These two assumptions are not always fulfilled in real-life traffic accidents.

Therefore, to address these issues, we assume that the characteristic vector X_i is not sufficient to capture all the differences between individuals, and we assume that there are other unobserved variables that can be represented by an additional random variable ε_i of the following form:

$$\lambda_i=\exp\left(X_i\beta+\varepsilon_i\right)$$

where, ε_i is a random variable that represents the various identification errors in λ_i due to the presence of uncontrollable influencing factors that cannot be controlled and therefore cannot be included in the model.

The marginal probability of an individual being involved in k_i accident is:

$$\int\Pr\left[k_i/X_i,\varepsilon_i\right]h\left(\varepsilon_i\right)d\varepsilon_i=\int\frac{e^{-\exp(X_i\beta+\varepsilon_i)}\exp\left(X_i\beta+\varepsilon_i\right)^{k_i}}{k_i!}h\left(\varepsilon_i\right)d\varepsilon_i$$

where, $h(\varepsilon_i)$ is the probability density function of ε_i, which is the general formula for the Poisson composite distribution.

Our special formula is written as:

$$\lambda_i = \exp\left(X_i\beta\right)\mu_i$$

Assuming that $\mu_i=exp(\varepsilon_i)$ follows the Gamma distribution with a probability density function

$$f\left(\mu\right) = \frac{\mu^{a-1}e^{-a\mu}a^a}{\Gamma\left(a\right)}$$

with a mean equals to 1 (The mean ε_i assumed to be equal to 0), and variance = $1/a$.

Therefore, mean λ_i is given by $exp(X_i\beta)$, and its variance is given by

$$\frac{1}{a}\exp\left(X_i\beta\right)^2$$

$$\begin{aligned}\Pr\left[N_i = k_i / X_i\right] &= \int \frac{e^{-\exp(X_i\beta)\mu_i}\left[\exp\left(X_i\beta\right)\mu_i\right]^{k_i}\mu_i^{a-1}a^{-a\mu_i}a^a}{k_i!\Gamma\left(a\right)}d\mu \\ &= \frac{\Gamma\left(k_i+a\right)}{k_i!\Gamma\left(a\right)}\left[\frac{\exp\left(X_i\beta\right)}{a}\right]^{k_i}\left[1+\frac{\exp\left(X_i\beta\right)}{a}\right]^{-(k_i+a)}\end{aligned} \tag{7}$$

which is the formula for a negative binomial distribution with parameters a and $\lambda_i=exp(X_i\beta)$, its mean and variance are respectively:

$$\begin{aligned}E\left(N_i\right) &= \exp\left(X_i\beta\right) \\ V\left(N_i\right) &= E\left(\lambda_i\right)\left[1+E\left(\lambda_i\right)V\left(\varepsilon_i\right)\right]\end{aligned} \tag{8}$$

Its variation is an increasing and convex transformation of the mean.

When analyzing the data collected on the insured, we find that a large number of them do not make any losses during the year, but the lack of losses may be actual, or it may be the result of not declaring the accident, i.e., a latent variable. These cases are handled using ZIP and ZINB models.

2.1.5 ZIP and ZINB models

Cragg (1971) developed various models to take into account the implicit variable mentioned above. In general, an event (e.g., buying an item, recognizing a loss...) may or may not occur. If the event does not materialize, this implicit variable takes a value of zero and is assumed to be an independent variable that takes positive values. The decision path is represented by the probit model, and the second event (number of incidents) is represented by the model defined in each case, so insureds with zero number of recorded losses (N=0) can be categorized into two categories:

A first category, none of which actually caused a loss.

Another group did not declare the accident because of its severity and to avoid the application of the penalty factor or to avoid the procedures for registering the accident.

This distinction is important for the insurer, as it can be conjectured that not declaring an accident for which the insured is responsible indicates that the risk is small, as the latter does not take the measures to declare it in order to preserve certain privileges, some of which we have already mentioned, but this does not mean that the insured has become a risk to the insurer.

The Poisson and negative binomial models do not allow us to distinguish between these two categories; however, the ZIP and ZINB models generate two separate models to be linked, which were developed by Greene (1994) and Lambert (1992) and assume only that zero and strictly positive values are generated by the same process.

Compared to the aforementioned Poisson and negative binomial models, here we assume that the random variable is the product of a binary law and Poisson's law (in the case of ZIP), or a negative binomial law (in the case of ZINB).

$$N = BN^{*} \tag{9}$$

The unobserved random variable is modeled by logistic regression in order to estimate the probability of being $k_i = 0$, i.e., specify the recorded incidents as zero for the insured i, $b_i = 0$, if the insured did not declare the accident and $b_i = 1$ in the reverse case. The random variable N^* follows a Poisson model (or negative binomial model) and is used to predict the value N for insureds who authorize the loss ($b_i = 1$). This equation estimates k_i expectation.

The models ZIP and ZINB have two parts, one part related to the counting model (for N^* which takes into account the number of losses in case the insured declares them) and another part related to the zero-amplification (why probit?) that explains the non-declaration.

More precisely, in the ZIP model, if we denote q_i the probability $b_i = 0$ (i.e., non-declaration) and λ_i, the Poisson's law parameter of the frequency of dependent losses as mentioned earlier to the explanatory variables (8), then the probability density of the variable is written as follows

$$P(N = 0 / X_i) = q_i + (1 - q_i)e^{-\lambda_i} \tag{10}$$

wherein $q_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$ and X_i, the matrix that describes the individual's characteristics. β, the vector of the coefficients for which an estimate is required, for non-zero k_i:

$$P(N = k_i / X_i) = (1 - q_i)e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \tag{11}$$

The conditional probability of the number of losses $b_i = 1$ is equal to the unconditional probability of the unobserved or unobserved variable k_i^*.

In the ZINB model, probability is given as

$$P(N = k_i / X_i) = q_i\left(1 - \min\{k_i, 1\}\right) + (1 - q_i)\frac{\Gamma(k_i + \upsilon)}{\Gamma(k_i + 1).\Gamma(\upsilon)}\left[\frac{\upsilon}{\upsilon + \lambda_i}\right]^{\upsilon}\left[\frac{\lambda_i}{\upsilon + \lambda_i}\right]^{k_i} \tag{12}$$

3. Methodology

3.1. Data Collection and Descriptive Analysis

The study population, which includes all auto insurance policies with yearly underwriting periods at agencies and insurance intermediaries connected to the Algerian Insurance Company's (SAA) regional directorate in Setif in 2023, as well as the losses reported during that time, must be defined before discussing the data collection method.

The process was carried out manually, and the data was only a sample of 520 units (insurance policies) underwritten in 2023; it was not exhaustive because it is not feasible. However, it was observed that:

- The process was done manually because there is no technology available to obtain the combined data digitally.
- The variable ageP, i.e., the age of the driver's license, is not expressed by the first date of issuance, but by the year of its renewal, and was therefore excluded from the study.
- The variable val-V, i.e., the price of the insured vehicle, is only known in the case of the theft and fire or comprehensive warranty and was also excluded from the study.

The study variables were defined as shown in the Table 1.

3.2. Modelling the Number of Accidents in Auto Insurance

3.2.1 Generalized linear model

In simple linear models, the dependent variable is expressed by a single explanatory variable, X. Whereas, in multiple models it is expressed by several explanatory variables X_i; in classical linear models, instead of expressing the dependent variable Y, it is expressed by its mathematical expectation $E(Y/X)$, whereas in generalized linear models (GLM) it is expressed by a cupulas, in order to interpret, we create a cupula between X and Y as we will detail below.

In 1972, Nelder & Wedderburn (1972) presented generalized linear models, which, like previous models, seek to determine the relationship between the explanatory and dependent variables (Compain, 2010). The diagram that follows provides a summary of how to design a generalized linear model (Nelder & Wedderburn, 1972).

Table 2 shows that choosing a y law from the exponential family is the first step in creating a generalized linear

model.

The following functions: $a()$, $b()$, $c()$, are determined by first independently estimating the dispersion parameter, which we later deem constant, in order to apply the generalized linear model as a general rule of thumb. Then choose the cupula. The parameters ($\beta_1, \ldots, \beta_p$) must then be estimated, in order to stabilize $\eta(X)$ and thus determine $\mu = g^{-1}(\eta(X))$ considered as a mean (expectation) of the model, which is finally stabilized θ which can be defined by $\theta = (b')^{-1}(\mu)$, which also allows for the calculation of the variance function $V(\mu)$, y variance. The aim of using cupula is to make the error variance more stable, and the simplest choice of cupula that simplifies calculations is choosing g that achieves $g = (b')^{-1}$.

Table 1. Econometric study variables

Type	Variable	Type Code	Explanation
Driver	Driver is the insured	type1	Takes the value 1 if the insured is the same as the driver, and 0 if not
		type2	Takes the value 1 if the insured is not the same as the driver, and 0 if they are
	Driver's age	ageC	Driver's age is a variable that takes on normalized values
	Driver's gendre	M	Takes the value 1 if the driver is male, and 0 if not
		F	Takes the value 1 if the driver is female, and 0 if not
Vehicle	Usage	Affaire	Takes a value of 1 if the use of the vehicle is business specific and 0 if it is not
		Fonctionnaire	Takes a value of 1 if the use of the vehicle is functional and 0 if it is not
		Commerce	Takes a value of 1 if the use of the vehicle is commercial and 0 if it is not
		auto-ecol, tax	Takes a value of 1 if the use of the vehicle is for driving instruction or a taxi and 0 if not
		TPM	Takes a value of 1 if the vehicle is used to transport goods and 0 if it is not
		TPV	Takes a value of 1 if the vehicle is used for passenger transport and 0 if it is not
		V. spécieux	Takes a value of 1 if the use of the vehicle is private use and 0 if it is not
	Vehicle's age	ageV	Vehicle age is a variable that takes normalized values (equal to 0 if the current year is the year of first use)
	Power	Puissance	Vehicle power
Guarantees	Guarantees types	Garan1	Takes a value of 1 if the selected Guarantee is Garan1 and 0 if it is not
		Garan2	Takes a value of 1 if the selected Guarantee is Garan2 and 0 if it is not
		Garan3	Takes a value of 1 if the selected Guarantee is Garan2 and 0 if it is not
Sinistre	Number of accidents	Nb	It takes normal values
	Amount of losses	Sinistre	A real positive variable
	Bonus-malus coefficient	b-m	It varies into 0.65 and 2

Source: Author's elaboration based on company's information

Table 2. GLM diagram

Stochastic Compound Explanation	Link	Regular Interpretive Compound
y follows an exponential law Its probability density function is given by $f_{\theta,\phi}(y) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\}$ So, we have $E(Y) = \mu = b'(\theta)$ $V(Y) = b''(\theta)a(\phi) = V(\mu)a(\phi)$	y expectation is symbolized by μ linked to $\eta(X)$ by a cupula symbolized $g(.)$, a monotonous and differentiable function, so it should be reversible $g(\mu) = \eta(X)$ a canonic cupula is a special function that satisfies $\mu = g^{-1}(\theta) \Leftrightarrow \theta = \eta(X)$	Let $x = (x_1, \ldots, x_p)$ be the number of Explained Variables observations, we define the linear expectation attached to the observation as $\eta(x) = \sum_{i=1}^{p} x_i \beta_i$ Parameters ($\beta_1, \ldots, \beta_p$) to be estimated are equivalent to θ

Source: Guillaume (2010)

The parameters of the generalized linear model are estimated using maximum likelihood. After determining the density function $f_{\theta,\phi}$, the logarithm of the likelihood can be written for the observations *i* by assuming that all observations have the same weight.

$$l_i = l(Y_i, \beta, \phi) = \ln\left(f_{\theta,\phi}\right) = \frac{\left(Y_i\theta - b(\theta)\right)}{\phi} + c(Y_i, \phi) \tag{13}$$

To find an estimate of $\hat{\theta}$ and $\hat{\phi}$, the logarithm of plausibility must be maximized; to do this we use iterative methods of maximization; we know that the estimator by maximum plausibility follows a normal asymptotic distribution and write:

$$\sqrt{n}\left(\beta - \beta\right) \xrightarrow{\mathcal{L}} N\left(0, \phi\left(X^T W X\right)^{-1}\right)$$

where,

$$W = diag\left(W_1, \ldots, W_n\right)$$

and

$$W_i = \frac{1}{V(\mu_i)} \times \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

So, we estimate *W* from $\hat{\beta}$ then we write $W^* = W(\hat{\beta})$, after that, we deduce:

$$\sqrt{n}\left(\beta_i - \beta_i\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma^2_{\beta_i}\right) \tag{14}$$

where,

$$\sigma^2_{\beta_i} = \left[\phi\left(X^T W^* X\right)^{-1}\right]_{ii} \text{ for } i \in 1; p$$

From this formula, we determine $q_{1-\frac{\alpha}{2}}$ the degree quantile $\left(1 - \frac{\alpha}{2}\right)$ of the natural law, and hence the confidence interval with (1-α) degree, for component number *i* of β is:

$$IC_\alpha(\beta_i) = \left[\beta_i - \frac{\sigma_{\beta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}; \beta_i + \frac{\sigma_{\beta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \tag{15}$$

The confidence interval for η_i and μ_i are given as follows:

$$IC_\alpha(\eta_i) = \left[\eta_i - \frac{\sigma_{\eta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}; \eta_i + \frac{\sigma_{\eta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \tag{16}$$

$$IC_\alpha(\mu_i) = \left[\mu_i - \frac{\partial \mu_i}{\partial \eta_i}\frac{\sigma_{\eta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}; \mu_i + \frac{\partial \mu_i}{\partial \eta_i}\frac{\sigma_{\eta_i}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \tag{17}$$

With

$$\eta_i = \phi X_i (X^T W^* X)^{-1} X_i^T \tag{18}$$

3.2.2 Generalized linear model fit and likelihood testing

In linear regression, we perform model fit tests from the sum of the residuals, while in generalized linear models we theoretically focus on Pearson and plausibility tests. To do this, we define the so-called model deviation as well as the Pearson statistic.

Estimating β with $\hat{\beta}$ using plausibility function allows us to obtain a maximization of plausibility for each observation, either by inference $\hat{\beta}$ or by implication $\hat{\mu}_i$:

$$\phi \times l\left(Y_i, \beta, \phi\right) = Y_i \theta - b\left(\theta\right) + cte \tag{19}$$

$$\phi \times l\left(Y_i, \beta, \phi\right) = Y_i \left(b'\right)^{-1}\left(\mu_i\right) - b\left(\left(b'\right)^{-1}\left(\mu_i\right)\right) + cte \tag{20}$$

If the model is good, the expectation $\hat{\mu}_i$ of the model corresponds to Y_i (where is the average Y_i under the hypothesis of multiple observations such as X = X_i). For a saturated model, we can calculate the logarithm of maximum likelihood as follows:

$$\phi \times l_{sature}\left(Y_i\right) = Y_i \left(b'\right)^{-1}\left(Y_i\right) - b\left(\left(b'\right)^{-1}\left(Y_i\right)\right) + cte \tag{21}$$

We define the model deviation, which measures the deviation between the plausibility of the model compared to the corresponding saturated model.

$$D = 2\phi \sum_{i=1}^{n}\left(l_{sature}\left(Y_i\right) - l\left(Y_i, \beta, \phi\right)\right) \geq 0 \tag{22}$$

We define standard deviation D^* as $D^* = D/\phi$, and we say that the model is more favorable when the deviation is close to zero, we use this result to test the reliability of the model, presented as a null hypothesis H_0: A model with p significant explanatory variables.

However, in practice, according to the hypothesis H_0, asymptotically D^* follows the Chi-squared law with a n-p degree of freedom. We say that a model is significant at α risk if the value $D^*$is less than or equal to the tabular value of the Chi-squared law at $\chi^2_{n-p}(1-\alpha)$.

However, this test is not effective in the case of binary variables that do not follow a Chi-squared distribution, in which case we resort to the Hosmer-Lemershow test, which is based on dividing $\hat{\mu}_i$ the rank ascending into categories g (often). The statistic used approximates the Chi-squared law with a degree of freedom g:

$$C^2 = \sum_{k=1}^{g} \frac{\left(\sum_{i=1}^{c_k} y_i - m_k^* \overline{\mu_k}\right)^2}{m_k^* \overline{\mu_k}\left(1 - \overline{\mu_k}\right)} \tag{23}$$

With $m_k^*(c_k)$ the number of heterogeneous observations in the class k, and $\overline{\mu_k} = \sum_{i=1}^{c_k} \frac{m_i}{m_k^*}\hat{\mu}_i$, where m_i is the number of observations in the class k.

We also know the Pearson statistic and it is often called the generalized Pearson's Chi-squared:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i + \mu_i\right)^2}{\mathrm{var}\left(y_i\right)} \tag{24}$$

If the distribution is normal and the link function is the same (identic), this statistic corresponds to the sum of squares of residuals (SCR).

To compare two models, we calculate the difference between their deviation $D = D_2 - D_1$, which follows a Chi-squared distribution with a degree of freedom $p_1 - p_2$ where p_1 and p_2 are the number of features in the first and second models, respectively.

There are also two other criteria for differentiating between models, AIC and BIC: The idea behind these two criteria is that the greater the plausibility of the model, the greater the logarithm of plausibility, which makes the model better. The relationships are given as follows:

$$AIC = -2\mathcal{L} + 2p$$
$$BIC = -2\mathcal{L} + p\log(n) \tag{25}$$

where, p is the number of estimated parameters, and n is the number of observations.

3.2.3 AFDM mixed analysis

The variables in this study are a combination of quantitative and qualitative variables, and we use what is known as AFDM (Pagès, 2004).

```
> par(mfrow=c(3,2))
> for(i in 7:10){
+ s.class(afdm0$li[,1:2],fac=TabAFDM[,i],clabel=0,cstar=0,cpoint=0.5,cellipse=0)
+ s.class(afdm0$li[,1:2],fac=TabAFDM[,i],cstar=0,cpoint=0,
+ col=rep("red",times=length(levels(TabAFDM[,i]))),add.plot=TRUE)}
> s.corcircle(afdm0$co[1:6,])
```

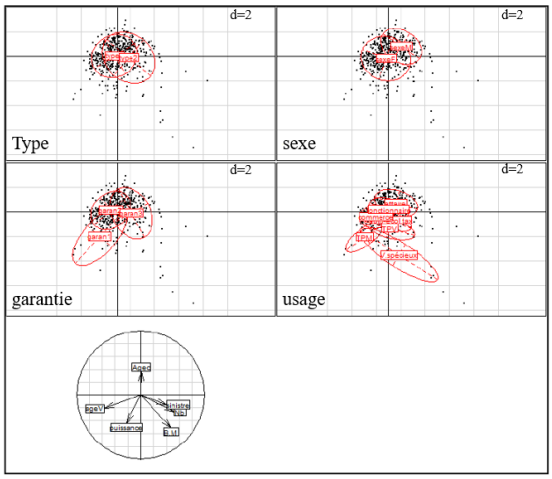

Figure 1. AFDM table
Source: Author's elaboration using R software

Therefore, we are now going to highlight a factor analysis of the mixed data using all the variables. Since the full study is quite large, we will not review all the results, but only focus on the most significant and important ones. Furthermore, we will not go into the details of results similar to those obtained in the multivariate component analysis example. We will begin by performing an AFDM on a table containing all the affected individuals, defined by all the explanatory variables. We will refer to the table in question as TabAFDM and list the names of the variables used below.

```
> names(TabAFDM)
[1] "Agec"      "B.M"       "ageV"      "puissance" "Nb"        "sinistre"
[7] "Type"      "sexe"      "garantie"  "usage"
```

The first six variables are quantitative variables, while the next four are qualitative variables.

Therefore, the output of the statistical program R will be through the following commands.

From the Figure 1, we can deduce several pieces of information. Firstly, we notice that the variables usage and garantie are correlated, so to improve the factor analysis and reduce noise, it is necessary to keep one of them, usage, as the dependent variable.

Next, to further express the variables, we categorize quantitative variables into qualitative variables (categories), which facilitates joint analyses. We rely on Charpentier (2013) and Ghali (2002) for our categorization.

The R outputs are as shown in Figure 2.

We may infer a great deal of information from the graph. First, we observe that the variables *b-m*, a*geV*, and *usage* are correlated. Because the agencies of the same organization are independent of one another, *b-m* can be removed from the study because it frequently fails to capture the true value of the reward and penalty coefficient, that is, it lacks memory. While *usage* clearly affects the number of losses, we keep the variable *ageV* since it has a higher effect on the amount of losses; the younger the automobile, the more guarantees and compensation.

The final representation of the factor analysis is as above, showing the independence between the variables.

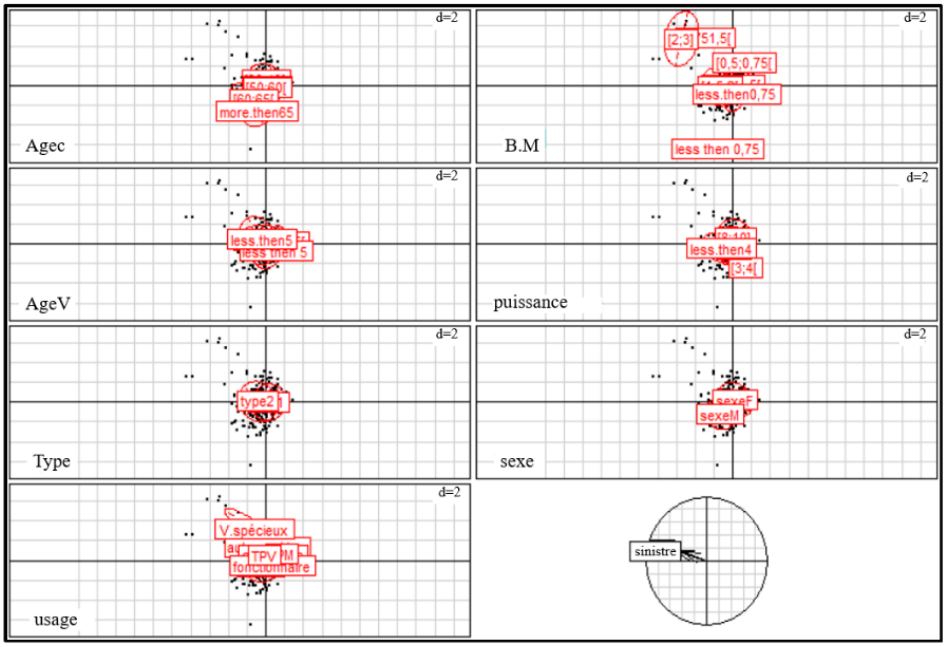

Figure 2. Categorizing quantitative variables into qualitative variables
Source: Author's elaboration using R software

4. Results and Discussion

The number of losses model in this article is the Poisson model because both the ZIP and ZINB models were not effective under the collected data. After entering the data and using the commands in the R program, an error showed up, and the commands used were as follows.

For the ZIP model, as shown in Figure 3.

```
> zeroinfl(Nb~Agec + ageV + B.M + usage + puissance + Type + sexe , data = data1, na.action=na.omit, dist = "
poisson")

Call:
zeroinfl(formula = Nb ~ Agec + ageV + B.M + usage + puissance + Type + sexe,
    data = data1, na.action = na.omit, dist = "poisson")

Count model coefficients (poisson with log link):
       (Intercept)               Agec               ageV               B.M
         -0.542747          -0.002169           0.009189          0.206973
usageauto-ecol,tax       usagecommerce  usagefonctionnaire          usageTPM
          1.308087           0.020250          -0.290925         -0.604063
          usageTPV     usageV.spécieux          puissance          Typetype2
          0.833599           1.257499          -0.137752          0.125969
         sexesexeM
          1.047288

Zero-inflation model coefficients (binomial with logit link):
       (Intercept)               Agec               ageV               B.M
          7.08869           -0.01726            0.11341          -5.24778
usageauto-ecol,tax       usagecommerce  usagefonctionnaire          usageTPM
          1.43988            0.09428           -3.63730         -0.76085
          usageTPV     usageV.spécieux          puissance          Typetype2
          1.17269            2.04057           -0.68543          -1.02077
         sexesexeM
          1.16306
```

Figure 3. The number of losses model for ZIP model
Source: Author's elaboration using R software

First section (Count model coefficients): The relationship between the variables and the number of non-zero accidents:

Usage auto-ecol, tax has a strong positive effect, meaning that using vehicles for commercial purposes increases the expected number of accidents.

SexeM indicates that men have a higher probability of the number of accidents compared to women.

Second section (zero-inflation model coefficients): Variables that affect the probability of zero-inflation:

B.M = -5.24778 shows a significant effect in reducing the likelihood of inflating the zero values, which means that this category explains most of the non-zero values. *Agec* and *PageV* have a relatively weak effect.

With AIC value = 738.2736.
For the ZINB model, as shown in Figure 4.

> zeroinfl(Nb~Agec + ageV + B.M + usage + puissance + Type + sexe , data = data1, na.action=na.omit, dist = "poisson")

```
Call:
zeroinfl(formula = Nb ~ Agec + ageV + B.M + usage + puissance + Type + sexe,
    data = data1, na.action = na.omit, dist = "negbin")

Count model coefficients (negbin with log link):
       (Intercept)               Agec               ageV               B.M
         -0.542882          -0.002168           0.009189          0.206986
usageauto-ecol,tax       usagecommerce  usagefonctionnaire          usageTPM
          1.308138           0.020259          -0.290905         -0.604030
          usageTPV     usageV.spécieux           puissance          Typetype2
          0.833593           1.257507          -0.137743          0.125987
         sexesexeM
          1.047303
Theta = 9907548.9473

Zero-inflation model coefficients (binomial with logit link):
       (Intercept)               Agec               ageV               B.M
           7.08828           -0.01725            0.11341          -5.24771
usageauto-ecol,tax       usagecommerce  usagefonctionnaire          usageTPM
           1.44003            0.09431           -3.63726          -0.76077
          usageTPV     usageV.spécieux           puissance          Typetype2
           1.17270            2.04065           -0.68541          -1.02074
         sexesexeM
           1.16315
```

Figure 4. The number of losses model for ZINB model
Source: Author's elaboration using R software

With AIC value = 740.2736.
For the negative binomial model, the outputs are as following Figure 5.

```
> glm.nb(formula = Nb~ Agec + ageV + B.M + garantie + usage + puissance + Type + sexe  , data = data1)

Call:  glm.nb(formula = Nb ~ Agec + ageV + B.M + usage + puissance +
    Type + sexe, data = data1, init.theta = 2.586174078, link = log)

Coefficients:
       (Intercept)               Agec               ageV               B.M
         -2.075939          -0.000423          -0.020861          0.853752
usageauto-ecol,tax       usagecommerce  usagefonctionnaire          usageTPM
          0.841840           0.012544           0.189162         -0.402607
          usageTPV     usageV.spécieux           puissance          Typetype2
          0.749382           0.545594           0.003964          0.260767
         sexesexeM
          0.790881

Degrees of Freedom: 519 Total (i.e. Null);  507 Residual
Null Deviance:      431.1
Residual Deviance: 383.5        AIC: 734.3
```

Figure 5. The number of losses using negative binomial model
Source: Author's elaboration using R software

To select only the variables that are most representative of the model, using the 'step' function as following Figure 6.

```
> glm.nb2=step(glm.nb,dir="backward")

Step:  AIC=724.59
Nb ~ ageV + B.M + sexe

       Df Deviance    AIC
<none>      380.39 724.59
- ageV  1   385.10 727.31
- sexe  1   394.05 736.25
- B.M   1   394.75 736.95
```

Figure 6. Representative variables of the model
Source: Author's elaboration using R software

To test the fit of the model and the explanatory variables, we used the function 'drop1' as following Figure 7.

```
> drop1(glm.nb2,test="Chi")

Model:
Nb ~ ageV + B.M + sexe
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      380.39 724.59
ageV    1   385.10 727.31  4.7179 0.0298498 *
B.M     1   394.75 736.95 14.3620 0.0001508 ***
sexe    1   394.05 736.25 13.6644 0.0002186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7. Fit test of the explanatory variables
Source: Author's elaboration using R software

The test shows that each of the variables *B.M.*, *sexe*, the variables of the model, are significant at the 0.1% risk degree.

We summarize the Poisson model as following Figure 8.

Created using the command in R:

```
>GLM.2 <- glm(Nb ~  Agec + ageV + B.M + usage + puissance + Type + sexe ,
 family=poisson(log), data=data1)
> summary(GLM.2)

Call:
glm(formula = Nb ~ Agec + ageV + B.M + usage + puissance + Type +
    sexe, family = poisson(log), data = data1)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.951809   0.586136  -3.330 0.000869 ***
Agec               -0.000554   0.006705  -0.083 0.934153
ageV               -0.021529   0.009458  -2.276 0.022830 *
B.M                 0.741954   0.258793   2.867 0.004144 **
usageauto-ecol,tax  0.827976   0.337478   2.453 0.014150 *
usagecommerce      -0.006214   0.215614  -0.029 0.977009
usagefonctionnaire  0.172130   0.327423   0.526 0.599088
usageTPM           -0.418815   1.016714  -0.412 0.680391
usageTPV            0.796602   0.367434   2.168 0.030158 *
usageV.spécieux     0.556184   0.395423   1.407 0.159559
puissance           0.007022   0.060154   0.117 0.907074
Typetype2           0.225783   0.179999   1.254 0.209713
sexesexeM           0.783935   0.183448   4.273 1.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 493.28  on 519  degrees of freedom
Residual deviance: 439.35  on 507  degrees of freedom
AIC: 735.77

Number of Fisher Scoring iterations: 6
```

Figure 8. Poisson model
Source: Author's elaboration using R software

To select the most representative variables, we use the 'step' function as following Figure 9.

```
> GLMA2=step(GLM.2,dir="backward")

Step:  AIC=730.13
Nb ~ ageV + B.M + sexe

       Df Deviance    AIC
<none>      451.71 730.13
- ageV  1   457.24 733.66
- sexe  1   467.79 744.21
- B.M   1   468.23 744.64
```

Figure 9. Representative variables selection
Source: Author's elaboration using R software

To test the fit of the model and the explanatory variables, we used the function 'drop1' as following Figure 10.

```
Model:
Nb ~ ageV + B.M + sexe
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      451.71 730.13
ageV    1   457.24 733.66  5.5274   0.01872 *
B.M     1   468.23 744.64 16.5135 4.831e-05 ***
sexe    1   467.79 744.21 16.0769 6.082e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10. Fit testing of the explanatory variables
Source: Author's elaboration using R software

In Figure 11, the test shows that each of the variables in the model is significant with a probability of error. Compare the two models.

```
> anova(GLM.2 , glm.nb2 , test="Chisq")
Analysis of Deviance Table

Model 1: Nb ~ Agec + ageV + B.M + usage + puissance + Type + sexe
Model 2: Nb ~ ageV + B.M + sexe
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       507     439.35
2       516     380.39 -9   58.966
```

Figure 11. Anova test
Source: Author's elaboration using R software

As can be shown, the probability value (Pr(Chi)) is not less than 0.05 because it does not appear. We get to the conclusion that there is little statistical support for the idea that the more intricate model (Model 2) offers a noticeably better match than the more straightforward model (Model 1). We can choose the Poisson model as a marginal model by comparing the AIC criterion for all models, which takes the values of 735 in the Poisson model, 734.3 in the negative binomial model, 740.2736 in the ZINB model, and 738.2736 in the ZIP model.

$$\begin{aligned}\log(\lambda) = &-2.075939 - 0.000423[\text{ageC}] - 0.020861[\text{ageV}] + 0.853752[\text{B.M }] \\ &+ 0.84184[\text{usage. auto-ecol, taxi}] + 0.012544[\text{usage. commerce}] \\ &+ 0.749382[\text{usage. TPV}] + 0.545594[\text{usage. V.sp?cieux }] \\ &+ 0.003964[\text{puissance}] + 0.260767[\text{Type.type2}] + 0.790881[\text{sexe.sexeM}]\end{aligned}$$

where, λ represents the average number of accidents $\lambda = E(N)$.

Explanation: For the variable *AgeC*, each one-year change in the age of the driver leads to an inverse change in the logarithm of the number of accidents with a value of 0.000423. Also, if the driver is male, this increases the value of the logarithm of the expected number of accidents by 0.790881.

Let us assume, for example, that the values for the independent variables are as follows in Table 3.

Table 3. An illustration of an application model

Variable	Value
Agec	30
ageV	5
B.M	1
Usage	"fonctionnaire"
puissance	7
Type	"type1"
sexe	"M"

Source: Author's elaboration

$$\begin{aligned}\log(\lambda) &= -2.075939 - 0.000423\times 30 - 0.020861\times 5 + 0.853752\times 1 \\ &\quad + 0.189162\times 1 + 0.003964\times 7 + 0.790881\times 1 \\ &= -0.331391\end{aligned}$$

The expected value of (number of incidents or cases) based on the input values is about 0.718.

5. Conclusions

This work highlights the need for a rigorous statistical technique, particularly GLM, in simulating auto accidents. This methodology can help insurance companies improve their risk management and underwriting processes. The results of this study can serve as a basis for more research and the development of policies aimed at reducing the number of traffic accidents in Algeria.

This study shows that the company's pricing mechanism is non-deterministic since other factors affect the price process. Based on the findings, the general model that was recommended as being most suitable for the business under study (SAA) is the Poisson model as a marginal model.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Cameron, A. C. & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *J. Appl. Econom.*, *1*(1), 29-53. https://doi.org/10.1002/jae.3950010104.

Charpentier, A. (2013). Actuariat avec R. *Freakonometrics*. https://doi.org/10.58079/ouoa.

Compain, H. (2010). *Analyse du risque de provisionnement non-vie dans le cadre de la réforme Solvabilité II* [Mastersthesis]. University of Paris Dauphine.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 829-844. https://doi.org/10.2307/1909582.

de Jong, P. & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755408.

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J. F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley.

Dionne, G. & Vanasse, C. (1989). A generalization of automobile insurance rating models: The negative binomial distribution with a regression component. *ASTIN Bull.*, *19*(2), 199-212. https://doi.org/10.2143/AST.19.2.2014909.

Ghali, O. N. (2002). Un modèle de tarification optimal pour l'assurance automobile dans le cadre d'un marché réglementé: Application à la Tunisie. *Assurances*, *69*(4), 603-654. https://doi.org/10.7202/1102480ar.

Greene, J. C. (1994). Qualitative program evaluation: Practice and promise. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 530-544). Sage Publications, Inc.

Guillaume, G. (2010). *Etude de la tarification et de la segmentation en assurance automobile* [Doctoralthesis]. Université Claude Bernard – Lyon 1.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica*, *52*(4), 909-938. https://doi.org/10.2307/1911191.

Lai, F. S. (2011). The accident risk measuring model for urban arterials. In *3rd International Conference on Road Safety and Simulation, Taiwan*.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14. https://doi.org/10.1080/00401706.1992.10485228.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2018). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood, Second Edition*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315119953.

Lemaire, J. (1985). *Automobile Insurance: Actuarial Models*. Kluwer-Nijhoff Publishing, Boston.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A*, *135*(3), 370-384. https://doi.org/10.2307/2344614.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Rev. Statist. Appl.*, *52*(4), 93-111.

Partrat, C. & Besson, J. L. (2005). *Assurance Non-Vie—Modélisation, Simulation*. Librairie Eyrolles. https://www.eyrolles.com/Entreprise/Livre/assurance-non-vie-9782717847062/

Vasechko, O. A., Grun-Réhomme, M., & Benlagha, N. (2009). Modélisation de la fréquence des sinistres en assurance automobile. *Bull. Fr. D'Actuariat*, *9*(18), 41-63.

Veysseyre, R. (2007). *Aide-Mémoire—Statistique et Probabilités Pour les Ingénieurs*. Dunod.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *J. Bus. Econ. Stat.*, *13*(4), 467-474. https://doi.org/10.2307/1392392.