Modeling of Motor Insurance Extreme Claims through Appropriate Statistical Distributions

V SELVAKUMAR¹, DIPAK KUMAR SATPATHI², P T V PRAVEEN KUMAR³, V V HARAGOPAL⁴

¹Department of Mathematics, Birla Institute of Technology and Science – Pilani, Hyderabad, India.
BHAVAN'S VIVEKANANDA COLLEGE OF SCIENCE, HUMANITIES, AND COMMERCE, HYDERABAD, INDIA.
Email: vskselva79@gmail.com

²Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: dipak@hyderabad.bits-pilani.ac.in

³Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: praveen@hyderabad.bits-pilani.ac.in

⁴Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: haragopal_vajjha@hyderabad.bits-pilani.ac.in

ABSTRACT

In the area of insurance, probability modeling has a wide variety of applications. In life insurance, the compensation sum is calculated in advance and may often be estimated using actuarial techniques, while in motor insurance, the claim amount is generally not known in advance. In the insurance business, improving actuarial risk control strategies is an essential technique for controlling insurance risk. Although an insurance company's risk assessment about its solvency is a complex and detailed problem, its solution begins with statistical modeling of individual claims' amounts. The main objective of this study is to determine an appropriate statistical distribution for third-party (TP) claim amount variable for 15,16,783 data claims of motor insurance data in India from 1985 to 2020. This article emphasizes the possible ways of obtaining a suitable probability distribution model that accurately explains insurance risks and how to use such a model for risk management purposes. For this reason, we have applied modern programming techniques, and statistical software implemented the methods provided based on data on premium amounts of third-party motor insurance claims. The fitted probability distribution will help the Insurance companies to predict the incidence of claims and assist insurers in handling unexpected insurance claim payments.

Keywords: Box-Whisker Plot; Skewed Distributions; Kolmogorov Smirnov test; Akaike Information Criteria; Bayesian Information Criteria

JEL Classification: C1, L2

Received: June 28, 2021

Accepted: September 30, 2021

Modelización de los Siniestros Extremos del Seguro del Automóvil mediante Distribuciones Estadísticas Adecuadas

V. SELVAKUMAR¹, DIPAK KUMAR SATPATHI², P.T.V. PRAVEEN KUMAR³, V. V. HARAGOPAL⁴

¹Department of Mathematics, Birla Institute of Technology and Science – Pilani, Hyderabad, India.
BHAVAN'S VIVEKANANDA COLLEGE OF SCIENCE, HUMANITIES, AND COMMERCE, HYDERABAD, INDIA.
Email: vskselva79@gmail.com

²Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: dipak@hyderabad.bits-pilani.ac.in

³Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: praveen@hyderabad.bits-pilani.ac.in

⁴Department of Mathematics, BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE – PILANI, HYDERABAD, INDIA.
Email: haragopal_vajjha@hyderabad.bits-pilani.ac.in

RESUMEN

En el ámbito de los seguros, la modelización probabilística tiene una gran variedad de aplicaciones. En los seguros de vida, el importe de la indemnización se calcula de antemano y a menudo puede estimarse mediante técnicas actuariales, mientras que en los seguros de automóviles, el importe del siniestro no suele conocerse de antemano. En el negocio de los seguros, la mejora de las estrategias actuariales de control del riesgo es una técnica esencial para controlar el riesgo de los seguros. Aunque la evaluación del riesgo de una compañía de seguros sobre su solvencia es un problema complejo y detallado, su solución comienza con la modelización estadística de los importes de los siniestros individuales. El objetivo principal de este estudio es determinar una distribución estadística adecuada para la variable del importe de los siniestros de terceros (TP) para 15.16.783 datos de siniestros de seguros de automóviles en la India desde 1985 hasta 2020. Este artículo hace hincapié en las posibles formas de obtener un modelo de distribución de probabilidad adecuado que explique con precisión los riesgos de los seguros y en cómo utilizar dicho modelo para la gestión de riesgos. Para ello, se han aplicado técnicas modernas de programación y un software estadístico ha implementado los métodos proporcionados a partir de los datos de los importes de las primas de los siniestros de los seguros de automóviles de terceros. La distribución de probabilidad ajustada ayudará a las compañías de seguros a predecir la incidencia de los siniestros y ayudará a las aseguradoras a gestionar los pagos de siniestros inesperados.

Palabras clave: Diagrama de caja y bigotes; Distribuciones asimétricas; Prueba de Kolmogorov Smirnov; Criterio de información de Akaike; Criterio de información bayesiano.

Clasificación JEL: C1, L2

Recibido: 28 de Junio de 2021

Aceptado: 30 de Septiembre de 2021

1. Introduction

Past claims have a significant link with future claims; this is why insurance firms find it challenging to apply past, or present claim amounts to predict premium expenditures in the areas of future claims. Further, claims that are voluminous have to find suitable statistical distribution for their large quantities. Then it is appropriate to see how well the statistical distribution holds up against their claims. This paper explores different ways to resolve these issues; an actuarial model is used in the sense of insurance claims data. The paper aims to produce a valid statistical claim amount distribution and investigate how well the distribution works. All kinds of insurance, including health, property, and industrial risk and liability, are general insurance. The insurance industry is a crucial sector for every country, regardless of its economic condition. It helps to minimize financial risk by sharing or pooling the liability among many people. As claims-driven as in insurance is its primary focus, the insurance industry employs many actuaries to understand and analyze the claims data. In motor insurance, large claim amounts infrequently occur, requiring high-tailed and highly skewed distributions like the Gamma, Weibull, Pareto, and lognormal to process and analyze the claim data (Philip J, 2016). As probability models for claims, we consider the likelihood of financial risks that people will incur and payout from non-life insurance providers due to insurable accidents as part of their contracts. Many efforts have been made in the literature by identifying a probabilistic model for the claims amount distribution in developed nations, most of which have focused on vehicle insurance model extensive coverage of claims. Empirical distribution functions may be essential for evaluating claims results, but claims should still be modeled to provide well-behaved probability distributions. These models have several purposes besides only predicting company revenue: providing premium loading, demonstrating the potential to create revenue estimates, reserves to maintain projected performance, and enabling deductibles to be calculated. As a result, this paper discusses the actuarial modeling measures to calculate a statistical fitting distribution for the claims data and verify if a given distribution suits the claims.

2. Related Work

The most important criterion in the auto insurance market is modeling the claim amounts based on appropriate distribution to decide high/low sums for future reserve planning. Usually, we find in the literature that the examined data follows either exponential distributions, Weibull, Gamma, Lognormal, or Pareto for well-fitted tails (V Packova and D Brebera, 2015), without specifying which distribution is the best among several distributions for the skewed data. The lognormal distribution will provide a suitable match for the loss data based on the literature on modeling the procedures (Marlin P, 1984). Most of the literature suggests that the lognormal distribution can yield accurate findings which can be used to draw inferences valid for decision-making in the non-life insurance business. When compared to a class of distributions composed of lognormal, exponential, gamma, and Weibull distributions, the lognormal distribution best fits the quantities claimed (Oyugi Margaret Achieng and NO, I, 2010). Based on (Krzysztof Bumeckia et al., 2000) report, property claim service providers offer indexes for damages originating from catastrophic events in the United States. In this study, the lognormal distribution provides a better fit than the Paretoian distribution. Another Study by (Omari C et al., 2018) proposed an empirical analysis amongst the continuous claims severity distributions fitted; the lognormal distribution is selected as a reasonably good distribution for modeling claims severity. While negative binomial and geometric distributions are chosen as the most appropriate distributions for the claims frequency compared to other standard discrete distributions.

On the other hand, (Akram Kohansal et al., 2019) found that the Kumaraswamy Marshal-Olkin Log-Logistic model is suitable for fitting Iranian Auto Insurance Claim data than the skew-normal model. (Eling Martin, 2012) found that the skew-student and skew-normal distributions are competitive compared to other parametric distributions based on property liability insurance data. Another research by (Mazviona and Chiduzza, 2013) recommends that the lognormal distribution provides the best fit in lower claims, while the Pareto distribution gives the best fit in the case of a large claim.

According to (Renshaw A.E, 1994), a generalized additive model for location, shape, and scale is used as a systematic modeling technique for evaluating claims processes in an employer's third-party liability insurance claim. Based on another study by (Edward. W Frees et al., 2008) proposed the negative binomial regression model that predicts the occurrence of a claim based on the factors, i.e., driver age, gender, a form of car, vehicle age, and no claim discount. The multinomial logit model often predicts the form of insurance claim dependent on injuries to third parties and damage to third party property. Also, Generalized beta (2), the third model for the severity variable, predicts claim amounts. Another study by (Daiane Aparecida Zuanetti et al., 2006) stated that claim size follows a Poisson distribution, and the claim amount will follow a lognormal distribution. In another study by (Yang et al., 2020), the Compound Poisson model applied to real and simulated data sets outperforms other models with accuracy and predictions. The mixed copula method makes it possible to rely on the number of claimants and their corresponding average size of claims by utilizing the Gaussian copula applied to the automobile insurance scheme portfolio. It suggests its supremacy over the classical Poisson compound model (Claudia Czado et al., 2012). The spatial modeling results significantly boost the claim frequency and claim size (Susanne G et al., 2007) by incorporating spatial dependence trends in a Poisson model for claim frequency and gamma distribution for average claim size models in non-life insurance. Another analysis by (V V Haragopal et al., 2017) found that the Generalized Pareto Distribution(GPD) provides the appropriate fit for the data, which contradicts the standard procedure of any insurance data fitting into the Lognormal Distribution of India. Also, (Zisheng and Chi X, 2006) stated that the GPD model predicts high quantiles and the maximum potential loss from medical insurance claims results. (Antonello Maruotti et al., 2018) adopt the maximum likelihood method to predict model parameters, provide the probability for all uncertain parameters for health care expenditure data. According to (Zubair Ahmad et al., 2019), the Weibull loss distribution, a three-parameter particular sub-model of the proposed family capable of modeling heavily tailed results. (Sattayatham and Talangtan, 2012) revealed that increasing the number of components(k) results in Finite mixture lognormal fitting is best for Motor insurance claims. This analysis focuses on this path and determines the effective distribution of India's motor insurance data for the extreme value modeling of data for a claim.

3. Objective of the Research

This research attempts to build a claim value model by adapting probability distributions based on historical sample data from chosen distribution families. To do this, we used historical claims data between 1985 and 2020. The study's key contributions are as follows:

- Modeling the claim paid amounts by fitting appropriate probability distributions
- Test if the selected distribution provides a good fitting of the data or not.

4. Research Design & Methodology

4.1 Data used for research:

The secondary data from different Insurance companies in India are collected from 1985 to 2020 for distinctive 36 years. The secondary data consists of 108 column variables and 15,16,783 row values (i.e., approximately one million five hundred sixteen thousand seven hundred eighty-three data points). We have studied the third-party claim amount variable for fitting an appropriate statistical distribution for 15,16,783 data claims using R Programming. We now discuss the techniques of model building, as explained below.

4.2 Methodology

The methodological portion focuses on modeling the claims with various distributions of likelihood from chosen families of distributions. An appropriate statistical pattern of recognition can be

diagnosed with a proper collection of data. Also, by systematically analyzing these motor insurance datasets, we can determine the distribution and ideal parameters for the dataset.

➤ **Modeling Process**

- **Identification of a model family of distributions:**
The assumption emerging from a particular distribution family is identified based on frequency distribution, graphical methods, knowledge about nature, and process decisions.
- **Parameter Selection:**
Determine the parameters based on the claim amount data and distributions using the maximum likelihood method and Method of Moments. Often, choose the parameters that decide a particular case of a distribution family.
- **The validity of the distribution:**
The chosen distribution's validity and parameters have a good fit for the data tested using statistical measures such as the Kolmogorov-Smirnov test, AIC, BIC, HQIC, and CAIC performance metrics.
- **Check the Model fit:**
Quantile-Quantile (Q-Q) plots are analyzed after validating the sample distribution to verify how well a theoretical distribution model suits. For a few distributions, Figure 2 displays the Q-Q Plots.

More effectively, claims data is usually skewed to the extreme right, so these claims can be modeled using a skewed distribution. Prior expertise and practice with curve fitting, computer tools, and exploratory data analysis were the deciding factors for choosing the appropriate distribution. The statistical software tool R Programming is used here to build the model. The following are some of the probability density functions (pdf) of the different probability distributions:

Gamma Distribution: The pdf of the random variable X is

$$f(x) = \frac{\lambda^n e^{-\lambda x} x^{\alpha-1}}{\Gamma(n)}, x > 0 \quad (1)$$

with parameters: shape $\alpha > 0$ and scale $\lambda > 0$

Pareto Distribution: The pdf of the random variable X is

$$f(x) = \left(\frac{\alpha \beta^{\alpha+1}}{x^{\alpha+1}} \right), x \geq \beta \quad (2)$$

with parameters: shape $\alpha > 0$ and scale $\beta > 0$

Lognormal Distribution: The pdf of the random variable X is

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2}, x > 0 \quad (3)$$

with parameters: location μ and scale $\lambda > 0$

Weibull Distribution: The pdf of the random variable X is

$$f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, x > 0 \quad (4)$$

with parameters: shape $\alpha > 0$ and scale $\beta > 0$

Log Logistic Distribution: The pdf of the random variable X is

$$f(x) = \left(\frac{\beta}{\alpha}\right) \frac{\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left(1 + \left(\frac{x}{\alpha}\right)^{\beta}\right)^2}, x > 0 \quad (5)$$

with parameters $\alpha < 0$ and $\beta > 0$

We have fitted above mentioned skewed distribution and also estimated the parameters using R Programming.

➤ Estimation

There are several methods to estimate the unknown population parameter like Method of Maximum Likelihood, Method of Moments, Method of Minimum Chi-Square, and Least Square Method. We have used the maximum likelihood estimator method to estimate the selected family distributions' unknown parameters among the four methods mentioned above. The Maximum Likelihood Estimation procedure states that an estimate of θ , $\hat{\theta}(x)$ within the acceptable range of θ , making the likelihood functions $L(\theta)$ as large as possible. i.e., for any admissible value $\hat{\theta}(x)$. $L(\hat{\theta}) \geq L(\theta)$. Thus $\hat{\theta}(x)$ is the solution of $\frac{\partial \log x}{\partial \theta} = 0$ and $\frac{\partial^2 \log x}{\partial \theta^2} < 0$ at $\theta = \hat{\theta}(x)$. Thus any solution $\hat{\theta}(x)$ of the equations which maximize $L(\theta)$.

➤ Validation of the model

The validity of various skewed distributions and their parameter derived is checked whether the fitted model is appropriate using various statistical tests. For all tests, the null hypotheses tested are

H_0 : The claim amount of data follows a specified distribution.

H_1 : The claim amount of data does not follow a specified distribution.

(i) Kolmogorov-Smirnov tests:

Under this test, the decision about H_0 is taken based on the empirical cumulative distributions of the claim amount data to the fitted cumulative distribution. If $F_n(x)$ and $F(x)$ are the observed distributions, respectively, the K-S statistics is

$$D_n = \max_{\text{overall } x} |F_n(x) - F(x)| \quad (6)$$

➤ Performance Criteria

We shall calculate four performance metrics such as AIC, BIC, HQIC, and CAIC consider such analytical measures to determine the goodness of fit among the applied distributions.

Akaike Information Criteria:

$$AIC = -2L_i + (2 * p) \quad (7)$$

Bayesian Information Criterion:

$$BIC = -2L_i + (p * \log(n)) \quad (8)$$

Hannan-Quinn Information Criterion:

$$HQIC = -2L_i + (2 * p * \log(\log(n))) \quad (9)$$

Consistent Akaike Information Criterion;

$$CAIC = -2L_i + \frac{2 * p}{(n - p - 1)} \quad (10)$$

Where L_i denotes a log-likelihood function of the distribution, P – number of model parameters of the distribution, n – total number of claims

➤ Diagnosis of Probability Models:

Quantile-Quantile plot is a helpful tool for evaluating distributional fit. Suppose we need to compare two distributions visually. In that case, we use the quantile plot, which visually compares two distributions by graphing the quantiles of one versus the other distribution quantiles.

Probability modeling in non-life insurance is challenging, and an excellent fit to the tail probability is crucial. This study has fitted five skewed distributions based on 15,16,783 data claims from different insurance companies of India. Also, we have considered the extreme claim amount and fitted different models distribution families, and checked the validity of the model to test whether the extreme claim amount follows a specified distribution or not.

5. Results and Discussion

The secondary sample data collected was analyzed using Mean, Median, Variance, Standard deviation, Skewness, and Kurtosis. Table 1 below shows the descriptive results of all claims paid

Table.1 Descriptive Statistics of Overall TP Claim Amount

Sample Size	Mean	S.D	Min	50% (Q2)	Max	Skewness	Kurtosis
1516783	154243	1273375	100	155508	30000000	8.30	112.21

From the above Table 1, we can observe the summary statistics of the overall third-party claim amount. The average claim paid was Rs. 154243, and the standard deviation was Rs. 1273375, and that of the skewness coefficient is 8.30, which indicates that the data is asymmetric. It suggests that the claim amount was positively skewed. Kurtosis has a value of 112.21; this measure's coefficient value indicates that the curve is Leptokurtic. We can see from the analytics that the claims' details have fewer claim amounts that occurred consistently. Whereas significant claims amounts infrequently happen, implying that a smaller claim sum has a higher relative frequency and a larger claim amount has a lower relative frequency.

Figure 1 Box Whisker Plot

From Fig. 1, It is observed that the Box-Whisker plot confirms the skew nature of the claims paid data. Also, Fig. 2 density plot indicates that the claim amount distribution is highly skewed with a long and heavy tail to the right.

Figure 2 Density trace of the overall Claim amount

The first step is to find a parameter of the particular statistical distribution to the claims data. In this case, the Maximum likelihood Estimation and Method of moments are used to determine the most likely probabilities for estimating the model parameter. Table 2 provides the parameters of five distributions fitted to the claims results. The estimated parameters of the distribution were used to find the log-likelihoods, AIC, and BIC values of the five distributions. Table 2 provides the parameters of five distributions fitted to the claims results.

Table 2 Parameter Estimation of Overall TP Claim Amount

Gamma	Weibull	Log-normal	Pareto	log-logistic
shape = 0.16949	shape = 0.6615632	Log mean = 12.032817	shape = 1.411091	shape = 1.158276
Scale = 0.0000003233	scale = 358737.825	Log std. dev. = 1.510262	scale = 271674.53	scale = 166070.094

The chosen distributions will be tested if they accurately describe the data or not by Akaike Information Criterion, Bayesian Information Criterion, log-like hood, and Goodness-of-fit tests. These findings indicate that the Weibull, Log-normal, Gamma, Pareto, and Log logistic distributions will suit the claims distribution. Now we are going to determine how well each distribution fits the results of our claims.

Table 3, summarises the findings of log-likelihood, Performance Criteria, and goodness of fit test values for the individual distributions. Based on the results, one can choose the model with the lowest AIC, BIC, HQIC, and CAIC values for modeling the overall TP claim data. In this case, the lognormal distribution had shown minimum performance criteria values. As a result, it was concluded that the lognormal distribution was the best-fitting distribution among the distributions chosen for the claims data, as it had the lowest AIC and BIC value. That is, of the five distributions considered, the lognormal distribution was calculated to be the most appropriate model to the unknown accurate distribution. It is advisable to consider the 'goodness of fit' until concluding how well the lognormal distribution matches the claims results. In this sequence, its Q-Q plots and P-P plots were plotted and interpreted as follows

Table 3 Performance Criterion of Overall TP Claim Amount

	Gamma	Weibull	Log Normal	Pareto	Log-logistic
Kolmogorov-Smirnov Statistic	0.00392510	0.00074006	0.00022942	0.00050864	0.00051061
log likelihood	-21833831	-21155062	-21028739	-21047447	-21041683
AIC	43667666	42310128	42057481	42094898	42083369
BIC	43667691	42310152	42057506	42094922	42083394
HAIC	43667672.6	42310134.6	42057488.6	42094904.6	42083376.6
CAIC	43667666	42310128	42057482	42094898	42083370

Figure 3 Q-Q and P-P plots for lognormal distribution Overall Claim amount

From fig.3, it was observed that the Q-Q plot, P-P plot, and Empirical & theoretical CDFs indicate the lognormal distribution was the best fitting distribution for overall TP Claim amounts. But, from Figure 4, the point falloffs fall closer to the line show that the usual distribution is most suitable for this data from the Quantile-Quantile plot. The diagonal line gives the optimal fit to the data with regard to other distributions. Unfortunately, all the skewed distributions have deviated from the extreme claim amount of TP around 5000000. It was inconclusive to determine which distribution fits the claim data because of the heavy tail.

For modeling the extreme claim amount of more than 5000000, 18241 distinct values between 5000001 and 30000000 claim amounts are taken. The effects of applying a two-parameter Pareto distribution to the data on X5000001 are seen in Tables 4 and 5. The appropriate way to view the fitted skewed distributions is to plot a histogram alongside approximate probability density functions for the data. Fig.5 represents the fitted frequency histogram.

Figure 4 QQ Plot of the overall Claim amount

Figure 5: Histogram, Density Plot, Box Plot for Extreme Claim amount

The equipped distribution's approximate parameters are described in Table 4. Table 5 contains the AIC, BIC, HAIC, CAIC, and goodness of fit test findings to determine if the two-parameter Pareto distribution matches the data adequately or not.

Table 4 Parameter Estimation of Extreme TP Claim Amount

Gamma	Weibull	Lognormal	Pareto	log logistic
shape = 4.022849	shape = 2.123524	Log mean = 15.9285843	shape = 1.98556	shape = 4.603061
scale = 0.0000004446	scale = 10256950	Log std. dev. = 0.3948172	scale = 5000000	scale = 7901207

After carefully and correctly conducting each phase of the actuarial modeling method, the subsequent analysis clearly shows that the Pareto distribution will better fit the extreme claim amounts compared to other distributions based on the AIC, BIC, HAIC, and CAIC values. Therefore, since the Insurance companies wish to model the motor insurance extreme claim amounts, the Pareto distribution will be the most suitable statistical distribution to generate accurate claim forecasts.

Table 5 Performance Criterion of Extreme TP Claim Amount

	Gamma	Weibull	Log Normal	Pareto	log-logistic
Kolmogorov-Smirnov Statistic	0.01818	0.01954	0.01102	0.01002	0.01085
log likelihood	-301558.4	-303393.9	-299484.2	-296282.8	-299232.9
AIC	603120.8	606791.9	598972.4	592569.5	598469.7
BIC	603136.4	606807.5	598988.0	592585.2	598485.4
HAIC	603125.9	606796.9	598977.5	592574.7	598474.9
CAIC	603120.80	606791.80	598972.40	592569.60	598469.80

After evaluating the Pareto distribution's goodness of fit graphically using Q-Q plots and mathematically using the AIC values, it is clear that the measures taken in the actuarial modeling phase are capable of producing accurate findings that can be used to draw useful inferences for decision making in the general insurance industry.

Figure 6 Q-Q Plot for Extreme Claim amount

The study's primary objective was to develop an appropriate statistical distribution for the third-party insurance claims data and determine how well this distribution matches the claims data to be used to model claim amounts. The analysis attempts to model the data's information to suit a specific distribution. As a result, an effort was made to determine an optimal statistical distribution that perfectly fits the insurance claims data by statistical computing and visual representation utilizing the R Programming.

From Table 3, we can observe that the lognormal distribution has the maximum log-likelihood function of -21028739, implying that it has a greater likelihood of fitting the claims data than the other statistical distributions such as log-logistic, Pareto, Weibull, and gamma distribution for overall TP claim data. Similarly, from Table 5, the Pareto distribution has the maximum log-likelihood function of

-296282.8 for the extreme claims, suggesting that it has a better chance of fitting the claims data than the other statistical distributions. Finally, the findings indicate that the column's log-likelihood values of all other distributions are far from adequate claims data models. The AIC, BIC, HAIC, and CAIC values are calculated to determine if the lognormal distribution suits the overall claims data well. In terms of AIC and BIC, the lognormal distribution has the lowest value of 42057481 and 42057506, meaning that it fits the overall TPclaims data well. Also, the Pareto distribution has the lowest value of 592569.5 and 592585.2 to provide extreme claims. Finally, in Figures 2 and 5, Q-Q plots for lognormal and Pareto distribution are plotted to visually confirm the goodness of fit test computed by the AIC fits well. As shown in Figure 5, the Pareto distribution's Q-Q plot offers an excellent fit to the extreme claims results, with nearly all points plotting on the line segment and just a few points fall near the line.

Many studies conveyed that different types of skewed distributions are appropriate for fitting their model based on the particular sample data in the literature. Our findings suggested that if the insurance claims amount data is huge (15,16,783 data claims), the lognormal distribution is an appropriate model for fitting the data and gives a much more accurate estimate for the third-party claim compared with other skewed distributions. Based on the analysis performed and the findings shown in Tables 3 & 5, it is determined that the lognormal distribution is the better fit for the overall insurance claims data, and the Pareto distribution fits well for extreme claims amount data.

6. Conclusions

This research demonstrated that the assumptions taken before analyzing insurance claims data could significantly impact the outcome. The objectives' beliefs resulted in selecting a family of distributions comprised of the Weibull, lognormal, Gamma, Pareto, and log-logistics. If the claim data becomes larger and larger, it showed that the lognormal distribution could have a good fit for the overall claim amounts. Also, the Pareto distribution was the best-fitting distribution among the distributions chosen for the extreme claim amounts. The empirical results demonstrate that the proposed distributions successfully model insurance losses. We expect that the approach can see broader use in actuarial sciences and other similar fields. Thus, increasing the sample distributions included in the study and ensuring the precision of the results. This analysis will help the Insurance companies to predict the incidence of claims and assist the companies in handling unexpected insurance claim payments.

References

1. Philip J. Boland (2016). Statistical methods in General Insurance. *International Association for Statistical Education*, ICOTS (7).
2. V Packova and D Brebera (2015). Loss Distribution in Insurance Risk Management. *Proceedings of the International Conference on Economics and Business Administration*, 17-22.
3. Marlin P (1984). Fitting the Lognormal Distribution to Loss Data Subject to Multiple Deductibles. *The Journal of Risk and Insurance*, 51(4), 687-701.
4. Achieng, O.M. and NO, I. (2010) Actuarial Modeling for Insurance Claims Severity in Motor Comprehensive Policy using Industrial Statistical Distributions. *International Congress of Actuaries*, Cape Town, Vol. 712, 7-12.
5. Krzysztof Burnecki, Grzegorz kukla and Rafal Weron(2000). Property Insurance Loss distribution. *Physica A: Statistical Mechanics and its Applications*, 287(1), 269-278.
6. Omari C, Nyambura S and Mwangi J (2018). Modeling the frequency and severity of Auto Insurance Claims using Statistical Distributions. *Journal of Mathematical Finance*, 8(1), 137-160.
7. Ramin Kazemi, Abdollah Jalilian and Akram Kohansal (2019). Fitting Skew Distributions to Iranian Auto Insurance Claim Data. *Applications and Applied Mathematics: An International Journal*, 12(2), 790-802.
8. Eling, Martin (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?. *Insurance: Mathematics and Economics*, Elsevier, 51(2): 239-248.

9. Batsirai Winmore Mazviona and Tafadzwa Chiduza (2012). The use of Statistical Distributions to Model Claims in Motor Insurance. *International Journal of Business, Economics, and Law*, 3(1), 44-57.
10. Renshaw A.E (1994). Modeling the claim process in the presence of covariates, *ASTIN Bulletin*, 24(2), 265-285.
11. Edward. W Frees and Emiliano A. Valdez (2008). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, 103:484,1457-1469.
12. Daiane Aparecida Zuanetti, Carlos A.R. Diniz and Jose Galvoa Leite (2006). A lognormal model for Insurance Claims Data. *RevStat- Statistical Journal*, 4(2),131-142.
13. Simon Fontaine, Yi Yang, Wei Qian, Yuwen Gu & Bo Fan (2020). A Unified Approach to Sparse Tweedie Modeling of Multisource Insurance Claim Data, *Technometrics*, 62:3, 339-356
14. Claudia Czado, Rainer Kastenmeier, Eike Christian Brechmann, and Aleksey Min (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 4(1), 278-305.
15. Susanne G and Claudia Czado (2007). Spatial modeling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007:3, 202-225.
16. V V Haragopal and K Navatha (2017). Fitting Distributions to Big Data: Example of Large Claim Insurance Data. *Visleshana, Big Data Analytics*, Vol.1.
17. Zisheng O, Chi, X (2006). Generalized Pareto distribution fit to medical insurance claims data. *Appl. Math. Chin. Univ*, 21, 21-29.
18. Antonio Punzo, Angelo Mazza & Antonello Maruotti (2018). Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions, *Journal of Applied Statistics*, 45:14, 2563-2584.
19. Zubair Ahmad, Eisa Mahmoudi and Gholamhossein G Hamedan (2019). A Family of Loss Distributions with an Application to the Vehicle Insurance Loss Data. *Pakistan Journal of Statistics and Operation Research*,15(3), 731-744.
20. Sattayatham, P and Talangtam, T (2012). Fitting of Finite Mixture Distributions to Motor Insurance Claims. *Journal of Mathematics and Statistics*, 8(1), 49-56.