Scientific Research Publishing

Applied Mathematics, 2025, 16(7), 584-592
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

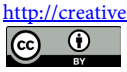
Comparison of Single and Composite Distributions in Modelling Auto Mobile Insurance Losses for Risk Measure Estimation

Williams Kumi[1,2*], Henry Otoo[1], Charles Kwofie[2], Sampson Takyi Appiah[2]

[1]Department of Mathematical Sciences, University of Mines and Technology, Tarkwa, Ghana
[2]Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana
Email: *williams.kumi@uenr.edu.gh

How to cite this paper: Kumi, W., Otoo, H., Kwofie, C. and Appiah, S.T. (2025) Comparison of Single and Composite Distributions in Modelling Auto Mobile Insurance Losses for Risk Measure Estimation. *Applied Mathematics*, **16**, 584-592.
https://doi.org/10.4236/am.2025.167032

Received: January 3, 2025
Accepted: July 25, 2025
Published: July 28, 2025



Abstract

Estimating risk associated with taking on random losses among various insurance policies is crucial, as it aids in obtaining the right balance between the reserve amount for paying indemnities and capital for investment to attract gains necessary to keep an insurance company in business. Fitting the right probabilistic distribution to insurance claims data is the first step in obtaining a good risk estimate for claim losses due to the complicated nature of claims data. Literature has shown that single distributions are incapable of holistically capturing the differences in claim losses due to the varying nature of small and large claims. Composite distributions, on the other hand, have shown tremendous promise in capturing the underlying dynamics that exist in claims data over some dispensations. To that end, in estimating risk measures associated with insurance losses in Ghana, this study first employs single distributions and then composite distributions to describe automobile insurance losses in Ghana for comparison. Two hundred and forty (240) composite distributions were fitted, and the top five were selected and presented taking into consideration the goodness of fit criterion: AIC, BIC, and Log-likelihood. Estimates of the composite distribution are computed using both numerical maximum likelihood estimation and general-purpose optimizers and numerical optimization techniques. The 16 single distributions that were combined to form the composite distributions were also separately fitted. For each of the top five selected single and composite distributions, Value at Risk (VaR) and Tail Value at Risk (TVaR) are estimated at 95% and 99% security levels. A comparison of the single and composite distributions showed that the composite distribution fitted better compared to the single distribution.

Keywords

Composite Distribution, Auto-Mobile Insurance, Claims, Risk Measures, Threshold, Mixing Weight

1. Introduction

One main function of an Actuary is to properly estimate risk, which can consequently be used in pricing and creating reserves. In estimating risk, it is, however, essential to determine the right probabilistic distribution of claims data. From the literature, single distributions such as Lognormal, Pareto, Weibull, and Gamma distributions have been widely used in modeling insurance claims. However, these distributions are not able to capture the extreme tail behavior of claims data, and this can lead to underestimation of the tail risk [1]. The failure of these traditional single distributions to accurately capture the dual nature (both small and large claims) of claims can lead to inadequate representation of tail behaviour essential for predicting insurance losses [2] [3].

Composite distributions, which combine multiple distributions, offer a flexible solution to this problem by accommodating the distinct characteristics of different segments of the data. The issue of appropriately modelling insurance claim data, particularly in the context of heavy-tailed data, has been a long-standing challenge in the field of actuarial science. In Ghana, this problem is especially pertinent due to the unique characteristics of the country's insurance market and the heavy-tailed nature of its claim data.

Risk management cannot be underrated in risky industries such as insurance, hence requiring extensive data modeling. It is therefore of great importance to understand and efficiently model claims distribution for robust risk management. Most often, there is a display of heavy tailness in claims data where extreme values depict low frequencies but are accompanied by high severities. These rare extreme events can have a detrimental and catastrophic impact on the insurance industry if not properly accounted for. A critical model is consequently required to predict these unforeseen circumstances and enable proper reserving to handle potential claims.

2. Methods

2.1. Composite Distribution

Here, we discuss in detail how to formulate composite distributions. Composite distributions join together two weighted distributions at a given threshold value. In statistical terms, let X be a random variable and let $f_1(.)$ be the pdf of the first distribution and the pdf of the second distribution be denoted by $f_2(.)$ of the same random variable. Let F_1 and F_2 be the corresponding cdf's of the random variables. The pdf of the composite model can then be expressed as:

$$f(\alpha_1,\alpha_2,\theta,\phi)=\begin{cases}\dfrac{1}{1+\phi}f_1^*(x/\alpha_1,\theta), & \text{if } 0<x\le\theta\\ \dfrac{\phi}{1+\phi}f_2^*(x/\alpha_2,\theta), & \text{if } 0<x<\infty\end{cases} \tag{2.0}$$

The continuity condition and the differentiability conditions are imposed at the threshold θ, such that in the limiting terms

$$f^+(\alpha_1,\alpha_2,\theta,\phi)=f^-(\alpha_1,\alpha_2,\theta,\phi) \tag{2.1}$$

$$f'(\alpha_1,\alpha_2,\theta,\phi)=f'(\alpha_1,\alpha_2,\theta,\phi) \tag{2.2}$$

where α_1 and α_2 are the parameters of the two pdfs on the two different intervals $(0,\theta]$ and (θ,∞), respectively. The differentiable and continuity conditions ensure θ and $\phi>0$ are defined as functions of the parameters α_1 and α_2. In addition, the mixing weights for the two density functions are given by $\dfrac{1}{1+\phi}$ and $\dfrac{\phi}{1+\phi}$. These mixing weights are defined as functions of α_1,α_2,θ. This can be written in closed form in terms of the cumulative density function as:

$$\phi=-\frac{\dfrac{\mathrm{d}\ln F_1(\theta/\alpha_1)}{\mathrm{d}\theta}}{\dfrac{\mathrm{d}\ln(1-F_2/\alpha_2)}{\mathrm{d}\theta}}=\frac{\dfrac{f_1(\theta/\alpha_1)}{F_1(\theta/\alpha_1)}}{\dfrac{f_2(x/\alpha_2)}{1-F_2(\theta/\alpha_2)}} \tag{2.3}$$

Substituting the expression for ϕ into the differentiability conditions simplifies to

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}\ln\left(\frac{f_1(\theta/\alpha_1)}{f_2(\theta/\alpha_2)}\right)=0\\ &\frac{f_1'(\theta/\alpha_1)}{f_1(\theta/\alpha_1)}=\frac{f_2'(\theta/\alpha_2)}{f_2(\theta/\alpha_2)}\end{aligned} \tag{2.4}$$

The functions $f_1^*(x/\alpha_1,\theta)$ and $f_2^*(x/\alpha_2,\theta)$ are truncated pdfs. In terms of their associated pdfs and cdfs is given by:

$$f_1^*(x/\alpha_1,\theta)=\frac{f_1(x/\alpha_1)}{F_1(\theta/\alpha_1)} \tag{2.5}$$

$$f_2^*(x/\alpha_2,\theta)=\frac{F_2(x/\alpha_2)}{1-F_2(\theta/\alpha_2)} \tag{2.6}$$

and also

$$F(\alpha_1,\alpha_2,\theta,\phi)=\begin{cases}\dfrac{1}{1+\phi}\dfrac{F_1(x/\alpha_1)}{F_1(\theta/\alpha_1)}, & \text{if } 0<x\le\theta\\ \dfrac{1}{1+\phi}\left[1+\phi\dfrac{F_2(x/\alpha_2)-F_2(\theta/\alpha_2)}{1-F_2(x/\alpha_2)}\right], & \text{if } 0<x<\infty\end{cases} \tag{2.7}$$

2.2. Model Selection Criteria

There are 16 loss distributions in the R-software package "actuar", which is gen-

erally accepted for modelling losses. We fitted 240 composite distributions from these loss distributions in "actua" by taking two distributions at a time, making 16C2 × 2 = 240 in total. The results of the top 5 composite distributions are presented based on the three goodness of fit criteria: AIC, BIC, and log-likelihood.

2.2.1. Value at Risk (VaR)

According to [4], Value at Risk (VaR) is a statistical measure that defines the worst expected loss over a specific time horizon under normal market conditions at a certain confidence level. It is commonly used in risk management to quantify the potential loss on an investment or portfolio. In simpler terms, VaR represents the maximum amount of money that could be lost on a portfolio within a set period of time with a specified level of confidence. The VaR at a 95% security level is calculated as follows:

$$\mathrm{VaR}_{\alpha}(X) = \pi_{\alpha} = F^{-1}(\alpha) \tag{2.8}$$

2.2.2. Tail Value at Risk

[5] and [6] defined the theoretical estimate for the TVaR of the random variable, X as:

$$\mathrm{TVaR}_{\alpha}(X) = \begin{cases} \dfrac{1}{1-\alpha}\left[\displaystyle\int_{\pi_{\alpha}}^{\theta} x f_1(x)\mathrm{d}x \frac{1}{F_1(\theta)} + \int_{\theta}^{\infty} x f_2(x)\mathrm{d}x \frac{1}{1-F_2(\theta)}\right], & \text{if } 0 < \alpha \le \dfrac{1}{1+\phi} \\ \dfrac{1}{1-\alpha}\dfrac{1}{1-F_2(\theta)}\left[\displaystyle\int_{\pi_{\alpha}}^{\infty} x f_2\right], & \text{if } \dfrac{1}{1+\phi} < \alpha < 1 \end{cases} \tag{2.9}$$

3. Results

3.1. Data and Its Source

Secondary data obtained from an insurance company was used for this work. It consists of one year of claim data, consisting of 11,892 data points.

3.2. Preliminary Analysis

The descriptive statistics of the data are presented in Table 1 below:

Table 1. Descriptive statistics of the comprehensive insurance data.

Statistic	Value (GH)
Mean	3884.2
1st Quartile	901.5
Median	2082.512 68
3rd Quartile	3961.2
Standard deviation	7542.157
Skewness	6.300 086 8
Minimum	20
Maximum	123,758.9

3.3. Fitting Single Distributions to Comprehensive Insurance Claims Data

Several single continuous distributions were fitted, but results from only the top five (5) are presented in Table 2 below. The table gives a summary of the parameter estimates and the AIC values. The AIC value for Lognormal was 5178.87, which was the lowest, indicating the best fit among the five distributions considered. This was closely followed by Gamma with an AIC of 5325.02 and Exponential distribution with an AIC of 5396.66. This explains the suitability of the Lognormal distribution as the appropriate model for the comprehensive data set.

Table 2. Summary of the top 5 single distribution and their parameters.

Distribution	Parameter Estimate	Goodness of Fit Criteria
Pareto	$\alpha = 2.861$ $X_m = 0.0581$	AIC = 5216 BIC = 5218.54 LL = −25582.43
Weibull	$\gamma = 0.7933$ $\alpha = 3.2971$	AIC = 5249.93 BIC = 5251.62 LL = −26247.47
Exponential	$\lambda = 0.2574$	AIC = 5396.66 BIC = 5397.01 LL = −26981.83
Gamma	$\beta = 0.74764$ $\alpha = 0.19253$	AIC = 5325.02 BIC = 5326.71 LL = −26624.01
Lognormal	$\mu = 0.55572$ $\sigma = 1.33259$	AIC = 5178.87 BIC = 5180.56 LL = −25892.94

3.4. Fitting Composite Distribution to Comprehensive Insurance Data

Table 3. The top five best composite distributions fitted to claims data.

Composite Distribution	Parameters (Head Distribution)	Parameters (Tail Distribution)	Goodness of fit
Gamma-Weibull	$\gamma = 1.11474$ $\beta = 0.04479$	$\gamma = 1.156204$ $\beta = 6.747606$	LL = 1661.04 AIC = 1330.08 BIC = 1359.46
Para logistic-Weibull	$\omega = 0.2574568$ $k = 0.03997525$	$\varepsilon = 3.422298$ $\upsilon = 796.9475$	LL = 1461.4 AIC = 1122.5 BIC = 1114.5
Inverse Paralogistic-Inverse Gaussian	$\omega = 1.19249095$ $k = 0.06319971$	$\mu = 5.766714$ $\sigma = 2.319205$	LL = 2057.2 AIC = 13299.39 BIC = 13291.39
Lognormal-Burr	$\mu = 6.163793$ $\sigma = 2.492292$	$\alpha = 0.01056385$ $c = 148.3444$ $k = 0.02816$	LL = 1540.6 AIC = 1091.2 BIC = 1081.2
Gamma-Invburr	$\alpha = 37.34696$ $\beta = 107.50197$	$\alpha = 0.44180161$ $\beta = 1.96312885$ $\gamma = 0.02609659$	LL = 1629.02 AIC = 1268.04 BIC = 1258.04

Table 3 above shows the results of the top five composite models out of the 240 composite models fitted to the data.

From the three criteria used in choosing the best candidate model, lognormal-burr was demonstrated to be the best candidate model for the data with an AIC of 1091.2, which is the least. This means that for the claims data, the body is best fitted with a lognormal distribution and the tail fitted with a burr distribution.

Table 4 below shows the threshold and mixing weights values for the top five composite distributions. The mixing weights which are given by $\frac{1}{1+\phi}$ and $\frac{\phi}{1+\phi}$ with $\phi > 0$ determines the appropriate segment of the data set which will be fitted to the body and tail distributions, respectively. For the best composite model, which is lognormal-Burr, 90.74% of the data points were fitted with the body distribution (Lognormal), whereas only 9.26% were fitted with the tail distribution. This clearly shows that the greater part of the losses was modelled with lognormal, whereas the remaining part was modelled by the burr distribution.

Table 4. Threshold and mixing weight values of the top five composite distributions.

Composite Distribution	Threshold (θ)	Weight Parameter (ϕ)
Gamma-Weibull	47.384 27	0.174 21
Paralogistic-Weibull	39.3100	0.0008
Inverse Paralogistic-Inverse Gaussian	25.9001	0.06086
Lognormal-Burr	36.5539	0.401 797 1
Gamma-Invburr	30.3393	354.3002

For the threshold and in a more general sense, given that X_i are the losses and θ is the threshold value, then $X_i < \theta$ are fitted by the body (head) distribution whereas $X_i \geq \theta$ are fitted by the tail distribution. The best composite distribution, lognormal-burr, estimated that losses up to GH 36,553.9 can be modelled by the head distribution (lognormal) and losses greater than GH 36,553.9 can be modelled by the tail distribution (Burr).

3.5. Estimation of the Risk Measures

Using the top five composite models, we now estimate the associated VaR and TVaR estimates at the security levels 95% and 99%. VaR is the worst possible loss an insurance company is likely to pay on any given trading day. The measure, when estimated, can be reliably used as an estimate for reserves. The right estimation of reserves is critical as it is necessary to ensure that only the right amount is set aside to pay claims, in order to invest any leftover money to generate investment income.

From Table 5, the Lognormal-Burr distribution estimated that at a security level of 99%, a typical insurance company can make a loss of GH 30116.00 or GH 33210.00, respectively, for VaR and TVaR on any given day. This amount is substantially huge, and as such, knowing this amount can help in properly reserving it in order to meet all obligations of paying claims.

Table 5. Risk measures of the top five composite models (in GH1000).

Composite Distribution	VaR_α		$TVaR_\alpha$	
	95%	99%	95%	99%
Gamma-Weibull	33.502	24.346	43.119	32.108
Para logistic-Weibull	31.108	25.034	39.211	31.022
Inverse Paralogistic-Inverse Gaussian	27.119	21.709	32.220	25.101
Lognormal-Burr	30.116	29.098	33.210	28.205
Gamma-Inverse burr	28.208	34.007	33.391	23.441

Now, given the differences in the VaR and TVaR values, it is essential to possibly find an average of the two risk measures to obtain an average risk estimate, which, to the best of our knowledge, may be more reliable. From the table, we realize that on average, from the two risk measures, a typical insurance company can pay claims of GH 31,663.00 on any given day. It is, however, important to note that this value is random.

3.6. Comparing Best Single Distribution to Best Composite Distribution

There are two folds in a comparison of the best single and best composite distribution. We compare their AIC values and probability plots (Q-Q plot) of the two distributions. From Table 2 and Table 3, the AIC values of lognormal and Lognormal-Burr distributions are respectively 5178.87 and 1091.2. A comparison of the values clearly shows that the composite Lognormal-Burr was a better fit than just using a lognormal distribution to fit the entire data.

Again, a comparison of the Q-Q plot of both the single distribution (Lognormal) in Figure 1 and the composite distribution (Lognormal-Burr) in Figure 2 shows that the composite distribution fitted the data better than the single distribution. This further supports the argument that composite distributions give a much better fit to claims data than a single distribution [5] [7] [8].

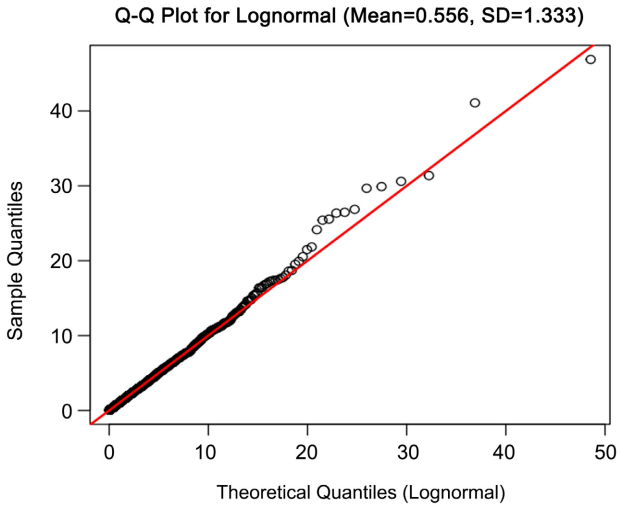


Figure 1. Q-Q plot of estimated Lognormal distribution.

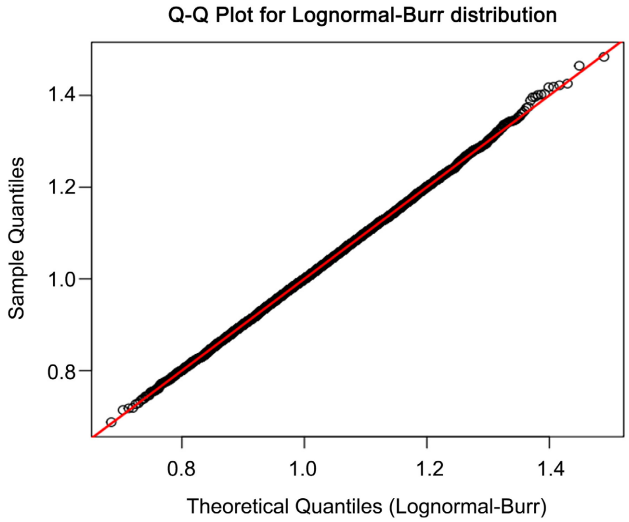


Figure 2. Q-Q plot of estimated Lognormal-Burr distribution.

4. Conclusions

In conclusion, in fitting probabilistic distributions for risk estimation to insurance claims data, composite models have demonstrated superiority over single distributions. The results of the study showed that Lognormal distribution was identified as the most appropriate single distribution for the Ghanaian comprehensive insurance claim data. Also, the best composite distribution for the data was the Lognormal-Burr distribution. A comparison of the best single and best composite distribution for the same data showed that the best composite distribution far outperforms the single distributions when comparing their goodness of fit criteria and Q-Q plots.

Due to the model's versatility in capturing extreme tail losses, Value at Risk (VaR) and Tail Value at Risk estimations were also estimated at the 95th and 99th percentile levels. The importance of using composite models for accurate risk estimation was expatiated in the study, which goes a long way to assisting insurance industries, particularly in markets such as the Ghanaian markets, where claims have heavy tails, since valuable insights for planning reserves and investment aspects of premiums collected by insurers were highlighted. The best fit composite distribution estimated a loss (VaR) of 30.116 (in thousands) at a 95% security level and 29.098 (in thousands) at a 99% security level. This loss amount is large and therefore needs careful planning of reserves to meet such obligations.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cooray, K. and Ananda, M. (2005) Modeling Actuarial Data with a Composite Lognormal-Pareto Model. *Scandinavian Actuarial Journal*, **2005**, 321-334. https://doi.org/10.1080/03461230510009763

[2] Mikosch, T. (2009) Non-Life Insurance Mathematics: An Introduction with Stochastic Processes. Springer.

[3] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2013) Loss Models: From Data to Decisions (vol. 715). John Wiley & Sons. https://doi.org/10.1002/9781118787106

[4] Jorion, P. (2001) Value at Risk: The New Benchmark for Managing Financial Risk. 2nd Edition, McGraw-Hill.

[5] Abu Bakar, S.A., Hamzah, N.A., Maghsoudi, M. and Nadarajah, S. (2015) Modeling Loss Data Using Composite Models. *Insurance*: *Mathematics and Economics*, **61**, 146-154. https://doi.org/10.1016/j.insmatheco.2014.08.008

[6] Grün, B. and Miljkovic, T. (2019) Extending Composite Loss Models Using a General Framework of Advanced Computational Tools. *Scandinavian Actuarial Journal*, **2019**, 642-660. https://doi.org/10.1080/03461238.2019.1596151

[7] Nadarajah, S. and Kwofie, C. (2022) Heavy Tailed Modeling of Automobile Claim Data from Ghana. *Journal of Computational and Applied Mathematics*, **405**, Article ID: 113947. https://doi.org/10.1016/j.cam.2021.113947

[8] Zhang, X. and Tang, W. (2013) Bayesian Estimation of Weibull-Pareto Composite Models in Health Insurance Claims. *Journal of Applied Statistics*, **40**, 2671-2672.