STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, July 2025, pp 1–19.
Published online in International Academic Press (www.IAPress.org)

The Synthetic Autoregressive Model for the Insurance Claims Payment Data: Modeling and Future Prediction

Heba Soltan Mohamed [1], Gauss M. Cordeiro [2], Haitham M. Yousof [3,*]

[1] *Department of Statistics and Quantitative Methods, Faculty of Business Administration, Horus University, Egypt*
[2] *Universidade Federal de Pernambuco, Departamento de Estatística, Brazil*
[3] *Department of Statistics, Mathematics and Insurance, Benha University, Egypt*

Abstract Time series play a vital role in predicting different types of claims payment applications. The future values of the expected claims are very important for the insurance companies for avoiding the big losses under uncertainty which may be produced from future claims. We define a new size-of-loss synthetic autoregressive model for the left skewed insurance claims. The new synthetic autoregressive model is assessed due to some simulations experiments. The optimal parameter is also artificially determined. The usefulness of the proposed model is proved by a real application.

Keywords Autoregressive Model; Claims payment; Cullen and Frey plot; Insurance Data; Residuals analysis; Statistical Forecasting.

AMS 2010 subject classifications 60E05; 62H05; 62E10; 62F10; 62F15; 62P05.

DOI: 10.19139/soic-2310-5070-1584

1. Introduction

Analyzing the insurance claims payment triangle from a U.K. Motor Non-Comprehensive account is crucial for understanding and predicting future claim liabilities. The data, spanning an origin period from 2007 to 2013, allows insurers to estimate outstanding claims reserves accurately. This analysis aids in assessing the financial impact of delayed claim settlements and helps maintain solvency. By modeling historical claims patterns, insurers can identify trends and seasonality, improving forecasting accuracy. It also supports better decision-making regarding premium setting and risk management strategies. Additionally, it enables companies to allocate capital efficiently by predicting cash flow needs for future claims. The analysis enhances regulatory compliance by ensuring adequate reserves are maintained. It also facilitates the evaluation of underwriting practices and claims handling efficiency over time.

In the insurance field, analyzing claims payment triangles is vital for effective reserving and risk assessment. It provides insights into the development patterns of claims over time, which is essential for accurate financial reporting. Understanding these patterns ensures that insurers remain financially stable by setting aside appropriate reserves. Furthermore, it supports the validation of pricing models and improves long-term strategic planning. This analysis is also instrumental in identifying anomalies or inefficiencies in claims processing, leading to operational improvements.

Box and Jenkins [5] developed autoregressive integrated moving average (ARIMA) models for prediction of time series data. The Box-Jenkins methodology designed for forecasting data is based on inputs from a specified time

[*] Correspondence to: Haitham M. Yousof (Email: haitham.yousof@fcom.bu.edu.eg). Department of Statistics, Mathematics and Insurance, Benha University, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

series. It can analyze several types of time series data for forecasting purposes. This methodology uses differences between time series data points to determine outcomes. It allows identifying trends using the autoregressive models, moving averages models and seasonal differencing to generate future forecasts. The ARIMA models are the main form of Box-Jenkins methodology. The two terms of these models are sometimes used interchangeably. For the time series analysis with forecasting and control, see Box et al. [6].

On the other hand, the actuarial literature became rich in researches based on the ARIMA models, see Cummins and Griepentrog [8] for forecasting automobile insurance paid claim costs using econometric and ARIMA models, Jang et al. [23] for analysing some medical insurance program for employees by ARIMA model, Venezian and Leng [37] for some applications of spectral and ARIMA analysis to combined-ratio patterns, Mohammadi and Rich [29] for the dynamics of unemployment insurance claims with an application of ARIMA model, Hafiz et al. [14] for projecting insurance penetration rate in Nigeria, and Kumar et al. [25] for forecasting motor insurance claim amount using ARIMA model.

The autoregressive (AR) and ARIMA models have attracted many authors in the field of applied mathematical modeling. The electricity price forecasting (Jakaša et al. [22]), modelling and forecasting of area, production, yield and total seeds of rice and wheat (Sahu et al. [33]), forecasting of wheat production (Iqbal et al. [21]), forecasting oil seeds prices in India (Darekar and Reddy [9]), forecasting wheat production in India (Nath et al. [31]), identification of paddy crop phenological parameters (Palakuru et al. [32]), and Shrahili et al. [36] for modeling the negatively skewed insurance claim-size asymmetric data using a new Chen model and the AR model.

The future insurance-claims forecasting is very important for insurance companies to avoid uncertainty about big losses that may be produced from future claims. Recently, Shrahili et al. [36] introduced a flexible claim-size Chen density for modeling asymmetric data (negative and positive) with different types of kurtosis (mesokurtic, leptokurtic and platykurtic). Since the insurance-claims data (Charpentier [6]) are a quarterly time series dataset, Shrahili et al. [36] analyzed these data using the AR model. A useful comparison is provided between the results of the Chen model and the autoregressive regression model. Many Chen densities were studied, see, for example, Ibrahim et al. [19] for a novel test statistic for right censored validity under a new Chen extension with some applications in reliability and medicine, Yousof et al. [40] for another Chen extension with characterizations and different estimation methods, and Korkmaz et al. [24] for a new unit-Chen model with associated quantile regression.

Following Shrahili et al. [36], we define a new size-of-loss synthetic autoregressive model (SAR) for the left skewed insurance claims datasets. The technique basically depends on exploring the insurance claims under all possible ARIMA models for selecting the best model. Then, this selection will depend on suitability for the insurance claims. The significance of the parameter model is statistically checked. The model with less number of significant parameter is preferable. The first step in developing a certain Box–Jenkins model for the time series insurance claims is determining whether the time series is stationary or not and whether there is any significant seasonality that requires to be modelled. After the Box–Jenkins model identification, the autoregressive model is chosen. The insurance claims are modeled using the SAR. Its adequacy is assessed through some simulation experiments. The optimal parameter is determined artificially.

The rest of the paper is organized as follows: Section 2 presents the SAR model along with its main statistical results. An assessment and application to historical insurance real data are addressed in Section 3. Finally, some concluding remakes are offered in Section 4.

2. The SAR model

The ARIMA is a class of statistical models that explains a given time series based on its own past values, i.e., its own lags and the lagged forecast errors, so that it can be used to forecast future values. The first step to construct a strong ARIMA model is to make the time series stationary. The SAR(p) is a linear regression model that uses its own lags as predictors, where p is the order of the SAR model. The linear regression models are adequate when the predictors are not correlated, and are independent of each other. We can explore and find out the required number of AR terms by inspecting the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots.

However, the PACF plots are more accurate than the ACF plots. For exploring the required number of AR terms, we shall present some simulated results.

Following Shrahili et al. [36], the new SAR model of order p (SAR(p)) can be expressed as

$$y_t = c + \vartheta_1 y_{t-1} + \vartheta_2 y_{t-2} ... + \vartheta_p y_{t-p} + \epsilon_t, \tag{1}$$

where ϵ_t is the white noise, c is a constant, the lagged values of y_t are the predictors, and $\vartheta_1, \ldots, \vartheta_p$ are the unknown parameters. We normally restrict autoregressive models to stationary data, where some constraints on the parameter values are required.

For the SAR(1) model:

- when $\vartheta_1 = 0$, then y_t is equivalent to white noise model which is ARIMA model with parameters (0,0,0);
- when $\vartheta_1 = 1$ and $c = 0$, then y_t is equivalent to a random walk model;
- when $\vartheta_1 = 1$ and $c = 0$, then y_t is equivalent to a random walk model with drift;
- when $\vartheta_1 < 0$, then y_t tends to oscillate around the mean.

Further, we can write

$$y_t = c + \vartheta_1 y_{t-1} + \epsilon_t \mid -1 < \vartheta_1 < 1, \, \forall \, t = 0, \pm 1, \pm 2, \ldots. \tag{2}$$

Here, the expected value of y_t is

$$\mathbb{E}\left(y_t\right) = 0 | t = 0, \pm 1, \pm 2, ..., \tag{3}$$

and its variance can be expressed as

$$\mathbb{V}\text{ar}\left(y_t\right) = \gamma\left(0\right) | t = 0, \pm 1, \pm 2, \ldots, \tag{4}$$

where

$$\gamma\left(0\right) = \Delta\left(\vartheta_1^2\right) \sigma_\epsilon^2 | \Delta\left(\vartheta_1^2\right) = \frac{1}{1 - \vartheta_1^2},$$

σ_ϵ^2 is the variance of the residuals, and the covariance $\mathbb{C}\text{ov}(y_t, y_{t-1})$ reduces to

$$\mathbb{C}\text{ov}\left(y_t, y_{t-1}\right) = \Delta\left(\vartheta_1^2\right) \vartheta_1 \sigma_\epsilon^2.$$

Analogously, the covariance $\mathbb{C}\text{ov}(y_t, y_{t-2})$ has the form

$$\mathbb{C}\text{ov}\left(y_t, y_{t-2}\right) = \Delta\left(\vartheta_1^2\right) \vartheta_1^2 \sigma_\epsilon^2.$$

For the SAR(2) model (or the ARIMA(2,0,0) model)

$$y_t = c + \vartheta_1 y_{t-1} + \vartheta_2 y_{t-2} + \epsilon_t|_{-1<\vartheta_1<1, \vartheta_2+\vartheta_1<1 \text{ and } \vartheta_2-\vartheta_1<1}. \tag{5}$$

The SAR in Equation (2) and some of its mathematical results will be used for statistical modeling of the claims payment data, and future prediction.

3. Assessment and application to historical insurance real data

Analyzing insurance claim amounts is a cornerstone of sound financial and actuarial management within the insurance industry. It serves as a critical mechanism for understanding the behavior of past claims and for projecting future liabilities with greater accuracy. By examining claim values across time, insurers can detect patterns, trends, and anomalies that reveal not only the frequency and severity of losses but also the operational performance of the claims process itself. This analysis enables insurers to estimate reserves for both reported and incurred-but-not-reported (IBNR) claims, which is vital for maintaining financial solvency and complying with regulatory frameworks such as Solvency II. Misestimating these reserves, either through under-reserving or over-reserving, can have significant consequences. Under-reserving may result in an inability to fulfill policyholder obligations, while over-reserving could lead to inefficient use of capital, reducing profitability and limiting investment opportunities.

From a financial standpoint, accurate claim analysis enhances the insurer's ability to manage risk exposure and make informed strategic decisions. It supports the development and refinement of pricing models by validating the assumptions upon which premiums are based. If historical claims data reveal that losses are consistently exceeding expectations, this may necessitate a review of underwriting guidelines, product pricing, or even the insurer's risk appetite. On the other hand, stable or improving claims performance can justify competitive pricing or expansion into new markets. Furthermore, detailed analysis of claim values aids in cash flow forecasting, helping insurers allocate capital efficiently to meet future claim obligations without holding excessive liquidity.

Additionally, analyzing claim amounts provides valuable insights into customer behavior, claim fraud risks, and process inefficiencies. It allows insurers to segment policyholders, identify high-risk profiles, and adjust coverage options accordingly. Operationally, it can expose systemic delays, data quality issues, or lapses in claims handling that may impact customer satisfaction and financial outcomes. Over time, continuous monitoring and analysis of claims data become a feedback loop that not only enhances internal processes but also contributes to better product design, customer service, and overall risk management. In summary, the analytical study of insurance claim amounts bridges the technical with the strategic—it is both a financial necessity and a competitive advantage in today's complex insurance landscape.

Estimating the parameters for all Box–Jenkins models involves numerically approximating the solutions of certain nonlinear equations. For this purpose, it is very common to use some statistical software like R. The two main approaches for fitting all these models are the nonlinear least squares and maximum likelihood methods. The second method is generally the most used in statistical literature, although it is too complicated for the full Box–Jenkins models, and then it is not included in this work.

The historical insurance real data are often reported in the form of a triangle presentation for showing the temporal development of claims overtime for each corresponding exposure (or origin) period. The exposure period could be considered as the year the insurance policy was earned, or the loss occurrence period. Clearly, the origin period does not have to be yearly. For example, it could be quarterly or monthly origin periods. The "claim age" or "claim lag" can be defined as the development period of an origin period. Data of individual policies are usually aggregated to homogeneous lines of business, division levels or perils. We analyze the insurance claims payment triangle from a U.K. Motor Non-Comprehensive account. For convenience, we set the origin period from 2007 to 2013 (Charpentier [6]). The insurance claims payment data frame presents the claims data in its typical form as it would be stored in a database. The first column holds the origin year (from 2007 to 2013), the second column is the development year, and the third column has the incremental payments. It is worth mentioning that these insurance claims data are firstly analyzed under a probability-based distribution.

We explore the insurance claims data. Exploring real data can be used either numerically or graphically or under both techniques. We consider many graphical techniques such as the skewness-kurtosis plot (or the Cullen and Frey plot) for exploring initial fits of theoretical distributions such as normal, uniform, exponential, logistic, beta, lognormal and Weibull. Bootstrapping is applied and also plotted for more accuracy. Cullen and Frey plot just compare distributions in the space of (the squared skewness, kurtosis), which is a good summary but still only a summary of the distribution properties.

Hence, many other graphical techniques are considered such as the "nonparametric Kernel density estimation (NKDE)" approach for exploring initial insurance claims density shape, the "Quantile-Quantile (Q-Q)" plot for exploring "normality" of the current data, the "total time on test (TTT)" plot for exploring the initial shape of the empirical hazard rate function (HRF), and the "box plot" for detecting the extreme claims. For more details about those graphical tools and their interpretation see Shrahili et al. [36], Mansour et al. [26], Mansour et al. [27], Mansour et al. [28], Hamedani et al. ([15],[16],[17],[18]), Nascimento et al. [30], Elgohari and Yousof [10], Goual et al. [13], Ibrahim et al. [20], Shehata et al. [34], Yousof et al. [39], Aboraya et al. [1], boraya et al. [2], El-Morshedy et al. [12], Elgohari and Yousof [11], Al-babtain et al. [3], Al-babtain et al. [4], Yadav et al. [38] and Shehata et al. [35]. For revealing the correlation between any two values of the signal changes as their separation

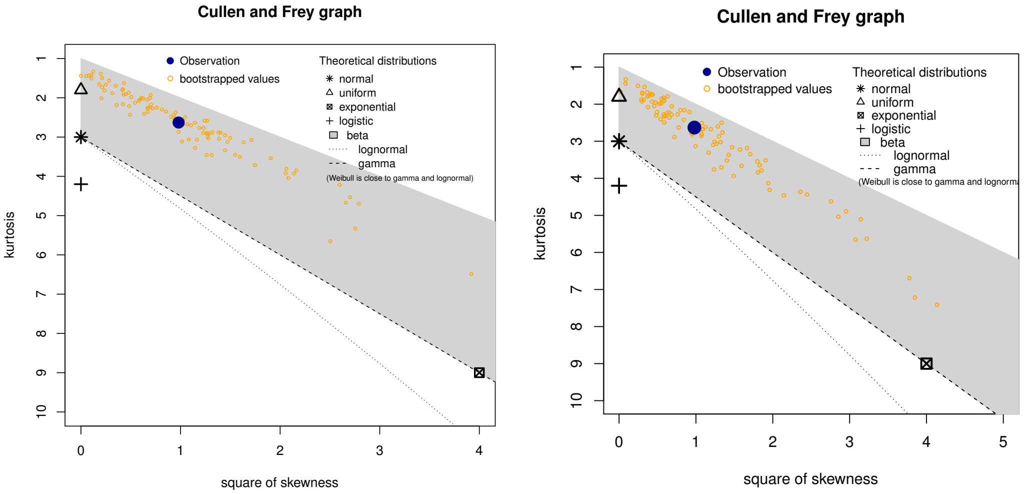

Figure 1. Cullen and Frey plot for the original insurance claims data (left) and the converted insurance claims data (right).

changes, we present the ACF. The theoretical ACF is a time domain measure of the stochastic process memory and does not reveal any information about the frequency content of the process. It provides some information about the distribution of hills and valleys across the surface with lag= $k = 1$.

Figure 1 displays the Cullen and Frey plot for the original insurance claims data (left plot) and the converted insurance claims data (right). Figure 2 and Figure 3 (top left plot) gives the NKDE plot for the original insurance claims data and the converted insurance claims data. Figure 2 and Figure 3 (top right plot) gives the Q-Q plot for the original insurance claims data and the converted insurance claims data. Figure 2 and Figure 3 (bottom left plot) displays the TTT plot for the original insurance claims data and the converted insurance claims data, and Figure 2 and Figure 3 (bottom right plot) gives the box plot for the original insurance claims data and the converted insurance claims data. Figure 4 shows the scattergrams (top plots), theoretical ACF (bottom left plot) and theoretical partial ACF (bottom right plot) for the original insurance claims data under lag= $k = 1$. Figure 5 displays the scattergrams (top plots), theoretical ACF (bottom left plot) and theoretical partial ACF (bottom right plot) for the converted insurance claims data under lag= $k = 1$. Figure 2 (top left plot) shows that the initial density for the the original insurance claims data is an asymmetric and bimodal function. Figure 3 (top left plot) indicates that the initial density for the the converted insurance claims data is an asymmetric function with left tail.

No extreme observations are spotted based on Figure 2 (bottom right plot) due to the original insurance claims data. Based one Figure 3 (bottom right plot), no extreme observations are spotted due to the converted insurance claims data. Further, Figure 2 (bottom left plot) shows that the HRF for the original insurance claims data is “monotonically increasing”, Figure 3 (bottom left plot) indicates that the HRF for the converted insurance claims data is “monotonically increasing”. Figure 4 (last plot) reveals that the first lag value are not statistically significant. Figure 5 (last plot) shows that the first lag value is statistically significant, whereas the other partial autocorrelations for all other lags are not statistically significant. This suggests a possible autoregressive (SAR(1)) model for these data. For the converted insurance claim’s payments data, skewness=-0.748278 (left-skewed data), kurtosis=2.788464<3 and dispersion index (Dis. Ix)=0.0708352 (underdispersed data). Based on these results, the SAR(1) model is suggested to explain the insurance claims data.

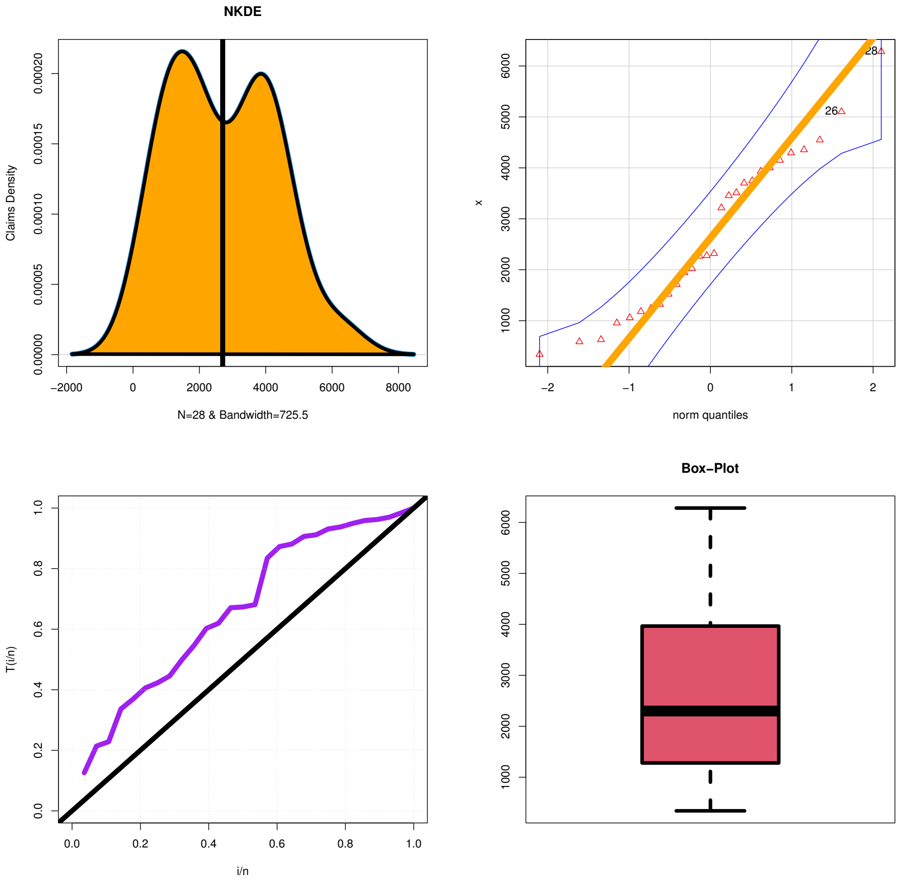

Figure 2. NKDE plot, Q-Q plot, TTT plot and box plot for the original insurance claims data.

Figures 6-13 display some artificial insurance claims data with generated ACF and partial ACF based on some positive and negative values of the parameter

$$\vartheta = (\pm 0.0001, \pm 0.01, \pm 0.1, \pm 0.15) \quad \text{and} \quad \vartheta = (\pm 0.20, \pm 0.30, \pm 0.40, \pm 0.50)\,.$$

For positive values of ϑ, the ACF exponentially decreases to 0 when the lag increases. For negative values of the parameter ϑ, the ACF also decays exponentially to 0 when the lag increases but the algebraic signs for the autocorrelations alternate between positive and negative. For positive values of ϑ, the partial ACF shuts off after the first lag since $\vartheta < 1$. For negative values of ϑ, the partial ACF shuts off after the first lag since $\vartheta < 1$.

Since the insurance claims data are quarterly time series, we will analyze the data using the SAR(1) model based on Figures 6-13. The SAR(1) model can be adopted for statistical forecasting of future insurance claims. However, the estimates of the parameter ϑ requires more efforts to be determined exactly. Table 1 provides $r_{[1]}$, $r_{[2]}$, $r_{[3]}$, $r_{[4]}$ and $\xi_{[11]}$ to find the exact value of ϑ. Note that

$$\xi_{[kk]} = 0, \ \forall\, k > 1.$$

The point prediction of the future values of $(Q_1)_{2014}, (Q_2)_{2014}, (Q_3)_{2014}$, $(Q_4)_{2014}$ and $(Q_4)_{2014}$ for the claim's payments in million. The future values $(Q_1)_{2014}, (Q_2)_{2014}, (Q_3)_{2014}$ and $(Q_4)_{2014}$ are very important for the insurance companies for avoiding the big losses under uncertainty which may be produced from the future claims.

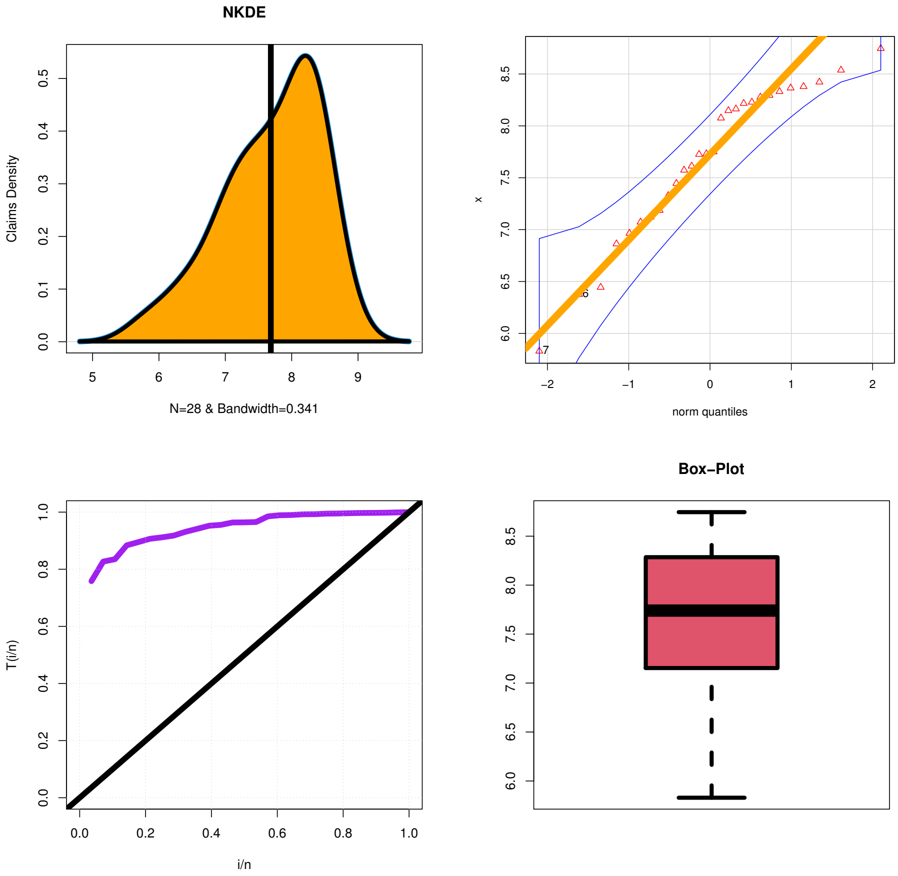

Figure 3. NKDE plot, Q-Q plot, TTT plot and box plot for the converted insurance claims data.

Tables 2-6 report the prediction errors (PEs), sum of prediction errors (SEs), mean errors of prediction (MEs), absolute percentage errors (APEs), sum of APEs (SAPEs), mean absolute percentage errors (MAPEs), absolute errors (AEs), sum of AEs, mean of AEs (MAEs), square errors (SEs), sum of square errors (SSEs), and mean square errors (MSEs) for $(Q_1)_{2014}$, $(Q_2)_{2014}$, $(Q_3)_{2014}$ and $(Q_4)_{2014}$. The results of $(Q_4)_{2013}$ are determined only for evaluating the SAR(1) model since the value of $(Q_1)_{2014}$ is already known. Based on Table 2-6, the SAR(1) model is suggested for determining the future values of $(Q_1)_{2014}$, $(Q_2)_{2014}$, $(Q_3)_{2014}$ and $(Q_4)_{2014}$ with $\vartheta = 0.5$. We have

- For $(Q_4)_{2013}|\vartheta = 0.5$ and $\xi_{[kk]} = 0 \,\forall\, k > 1:$
  PE$_{(Q_4)_{2013}}$=| PE$_{(Q_4)_{2013}}$|=0.387899, APE$(Q_4)_{2013}$=0.0460554, PE$^2_{(Q_4)_{2013}}$=0.1504656;
- For $(Q_1)_{2014}|\vartheta = 0.5$ and $\xi_{[kk]} = 0 \,\forall\, k > 1:$
  PE$_{(Q_1)_{2014}}$=|PE$_{(Q_1)_{2014}}$|=0.581849, APE$(Q_4)_{2013}$=0.06908316, PE$^2_{(Q_1)_{2014}}$=0.3385483;
- For $(Q_2)_{2014}|\vartheta = 0.5$ and $\xi_{[kk]} = 0 \,\forall\, k > 1:$
  PE$_{(Q_2)_{2014}}$=|PE$_{(Q_2)_{2014}}$|=0.678823, APE$(Q_2)_{2013}$=0.08059692 , PE$^2_{(Q_2)_{2014}}$=0.4608007;
- For $(Q_3)_{2014}|\vartheta = 0.5$ and $\xi_{[kk]} = 0 \,\forall\, k > 1:$
  PE$_{(Q_3)_{2014}}$=|PE$_{(Q_3)_{2014}}$|=0.727311, APE$(Q_2)_{2013}$=0.08635392 , PE$^2_{(Q_3)_{2014}}$=0.5289813;
- For $(Q_4)_{2014}|\vartheta = 0.5$ and $\xi_{[kk]} = 0 \,\forall\, k > 1:$ PE$_{(Q_4)_{2014}}$=|PE$_{(Q_4)_{2014}}$|=0.751554, APE$(Q_4)_{2013}$=0.08923230, and PE$^2_{(Q_4)_{2014}}$=0.5648334.

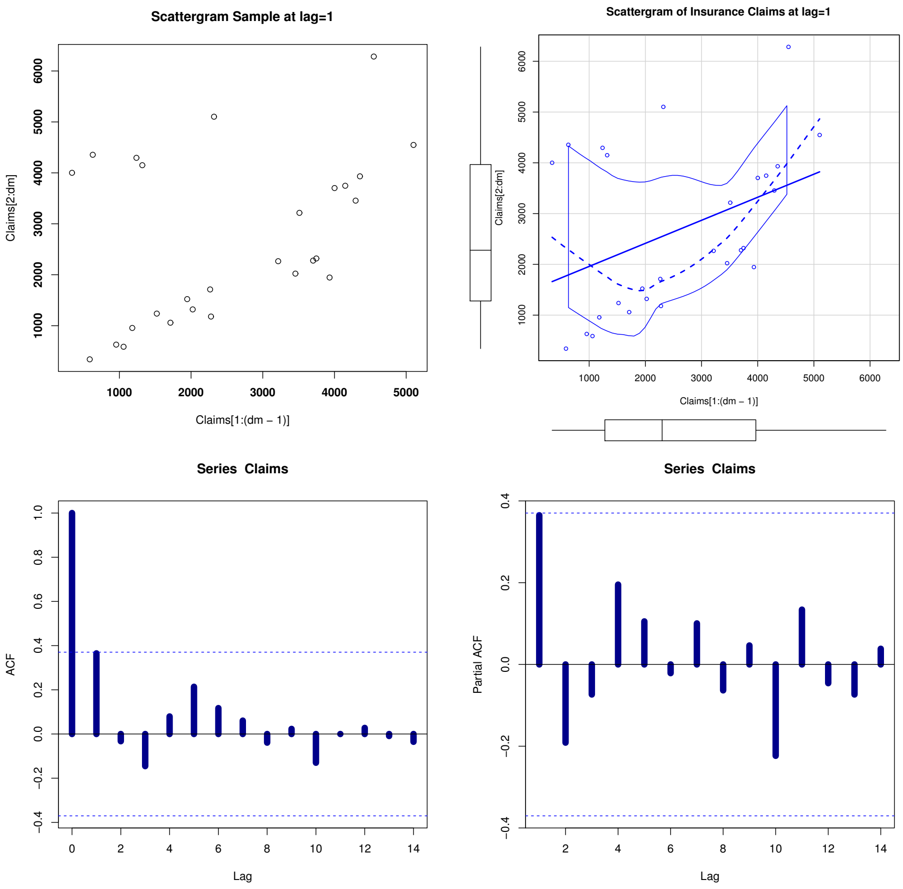

Figure 4. Scattergrams and autocorrelation function for the original insurance claims data.

It is also noted that MSEs<(MEs=MAEs)<MAPE for all cases (see Table 2). Figure 14 displays the scattergrams for the forecasting residuals under ϑ^+ for the future values $(Q_4)_{2013}$, $(Q_1)_{2014}$, $(Q_2)_{2014}$, $(Q_3)_{2014}$ and $(Q_4)_{2014}$. Figure 15 displays the scattergrams for the forecasting residuals under ϑ^- for the future values $(Q_4)_{2013}$, $(Q_1)_{2014}$, $(Q_2)_{2014}$, $(Q_3)_{2014}$ and $(Q_4)_{2014}$. The MSEs are preferable than the values of MEs if we have some negative errors.

Fortunately, we have no negative errors. The AEs, sum of AEs, MAEs are equal to the Es, SEs, MEs, respectively, since all errors are positive in this case. Hence, the results of absolute errors, sum of absolute errors, and mean absolute errors are omitted.

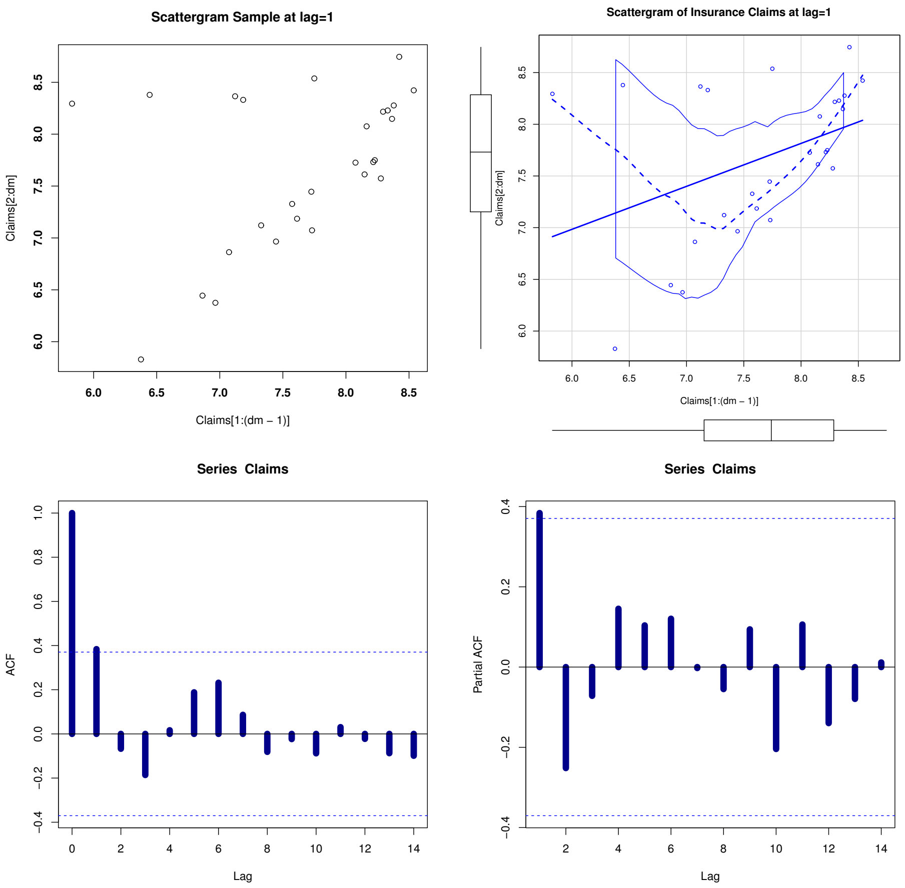

Figure 5. Scattergrams and autocorrelation function for insurance claims data.

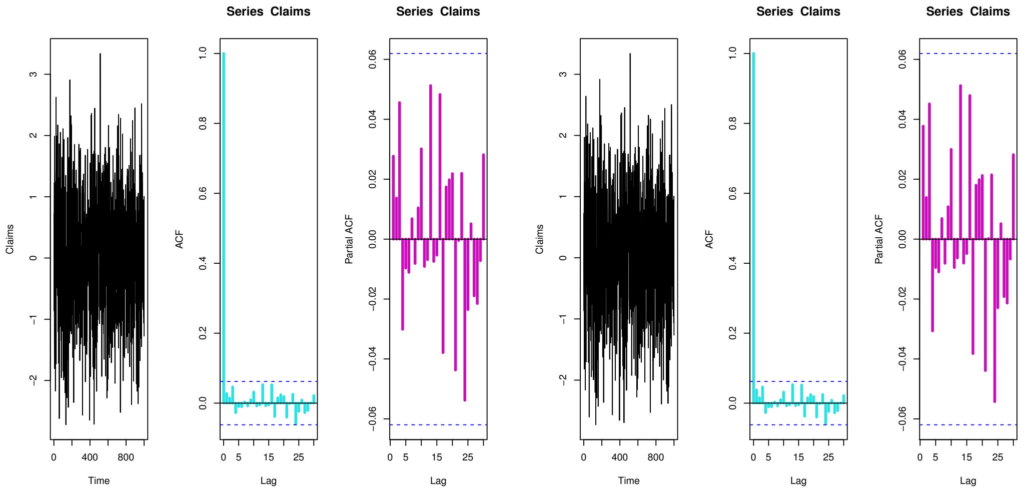

Figure 6. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.0001, -0.0001$

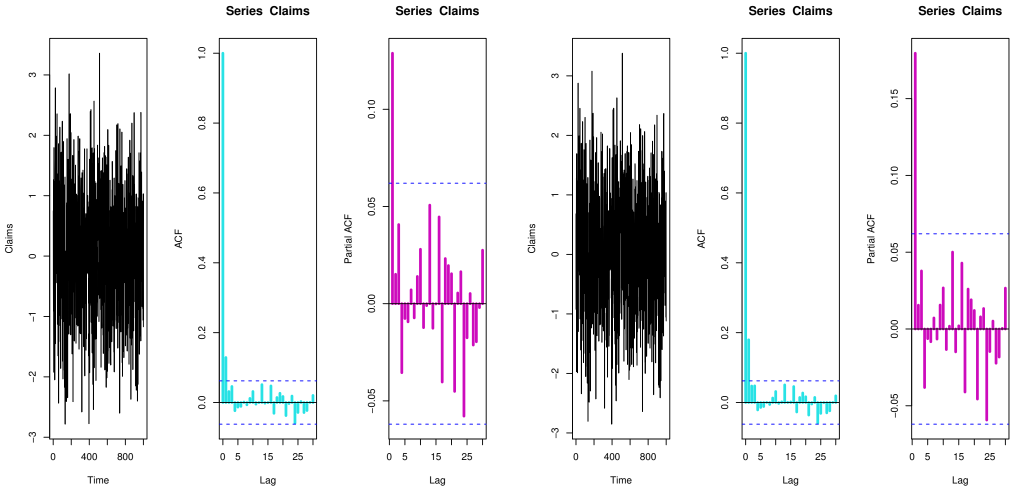

Figure 7. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.01, -0.01$.

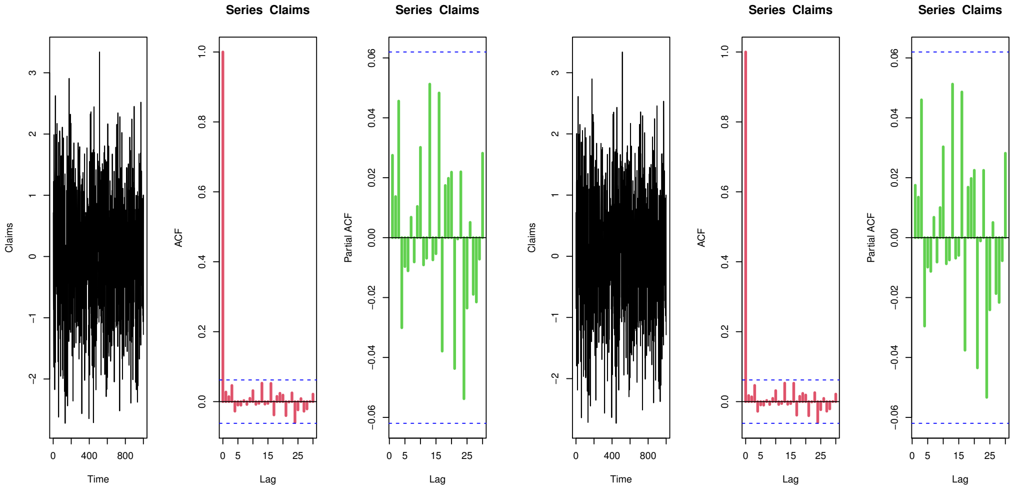

Figure 8. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.1, -0.1$.

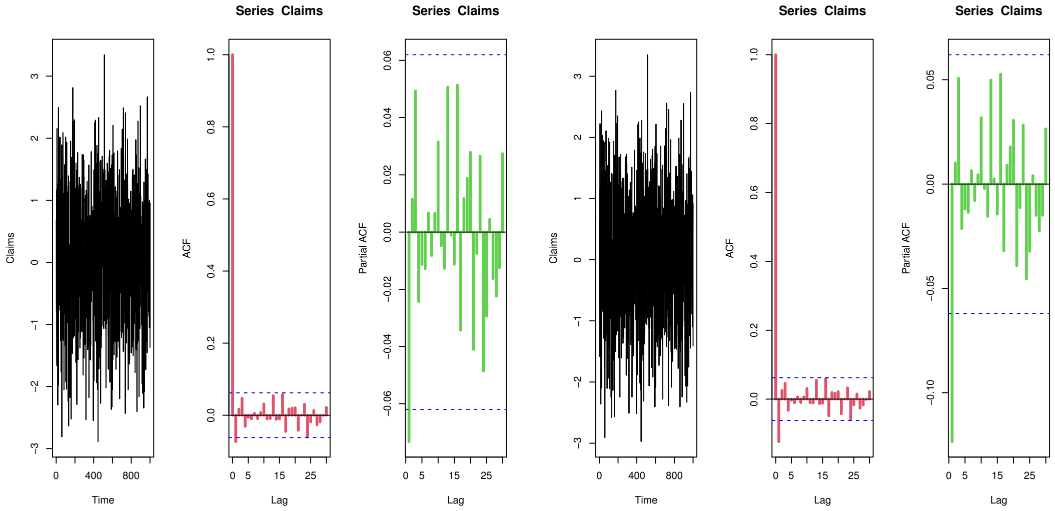

Figure 9. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.15, -0.15$.

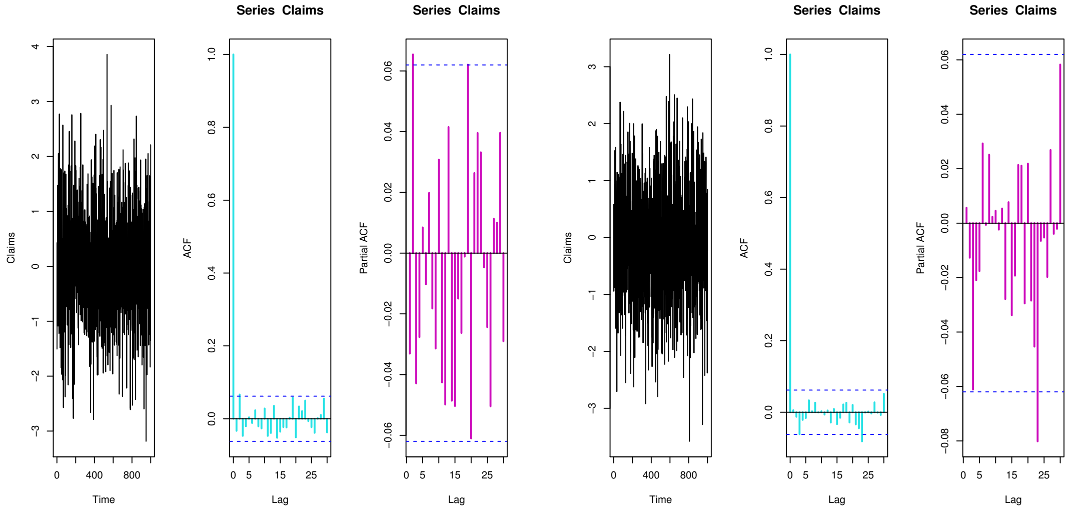

Figure 10. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.2, -0.2$.

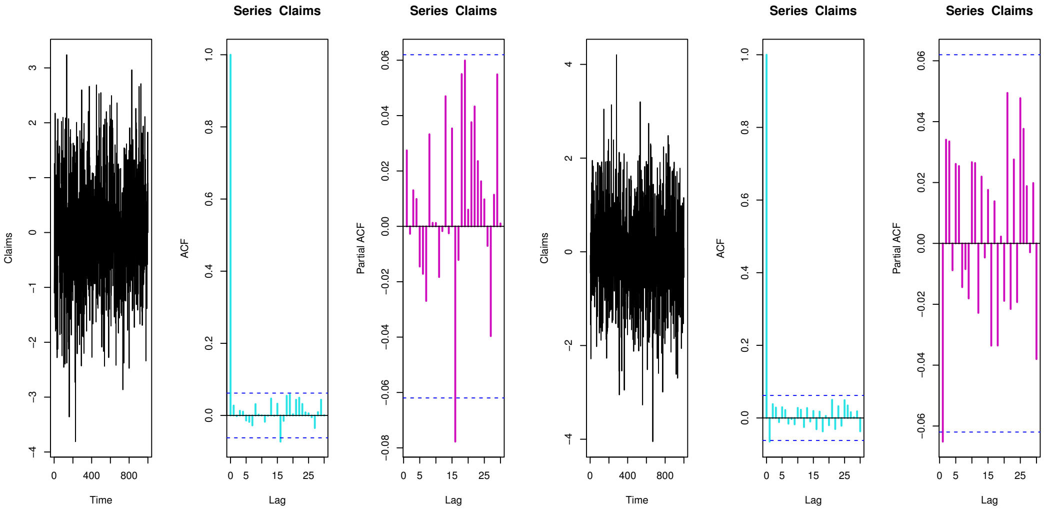

Figure 11. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.3, -0.3$.

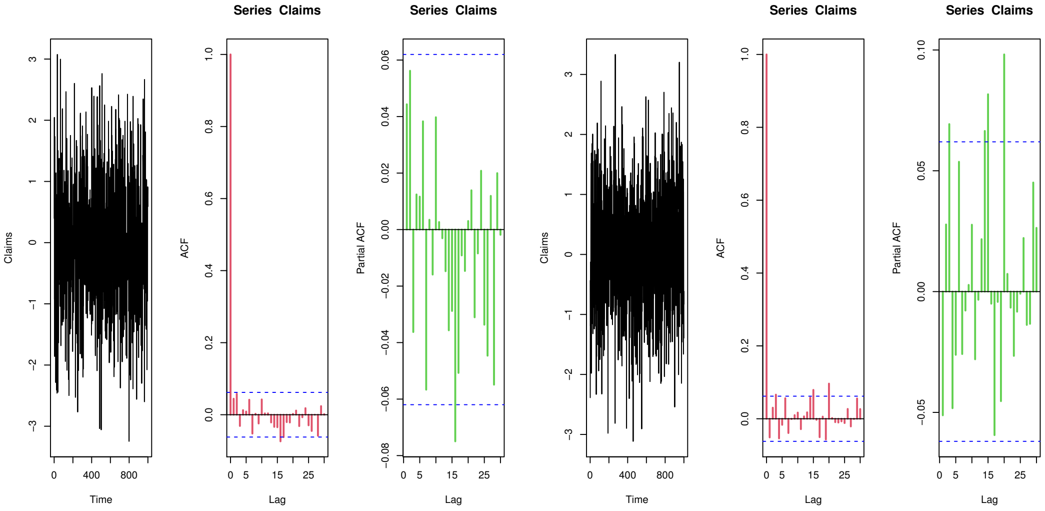

Figure 12. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.4, -0.4$.

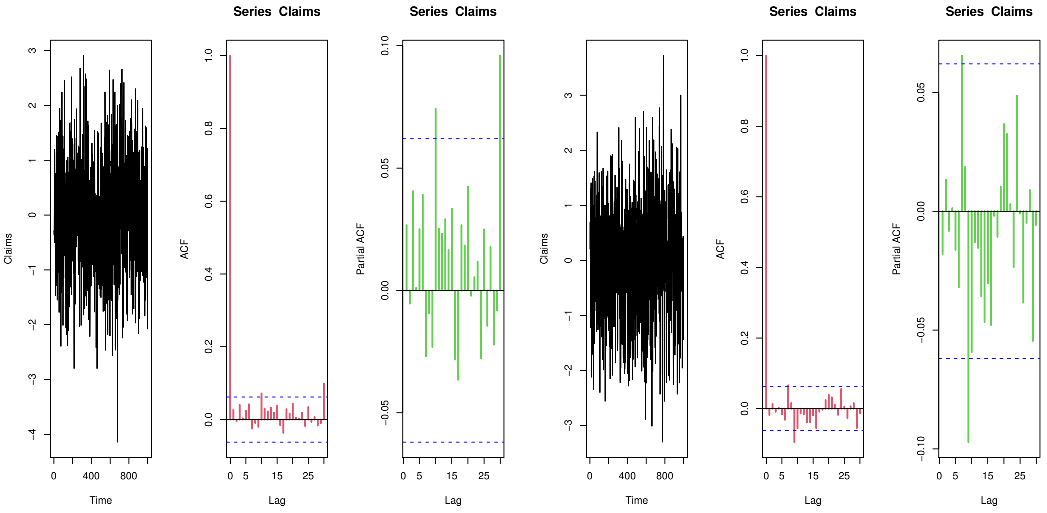

Figure 13. Artificial insurance claims data with ACF and partial ACF for $\vartheta = 0.5, -0.5$.

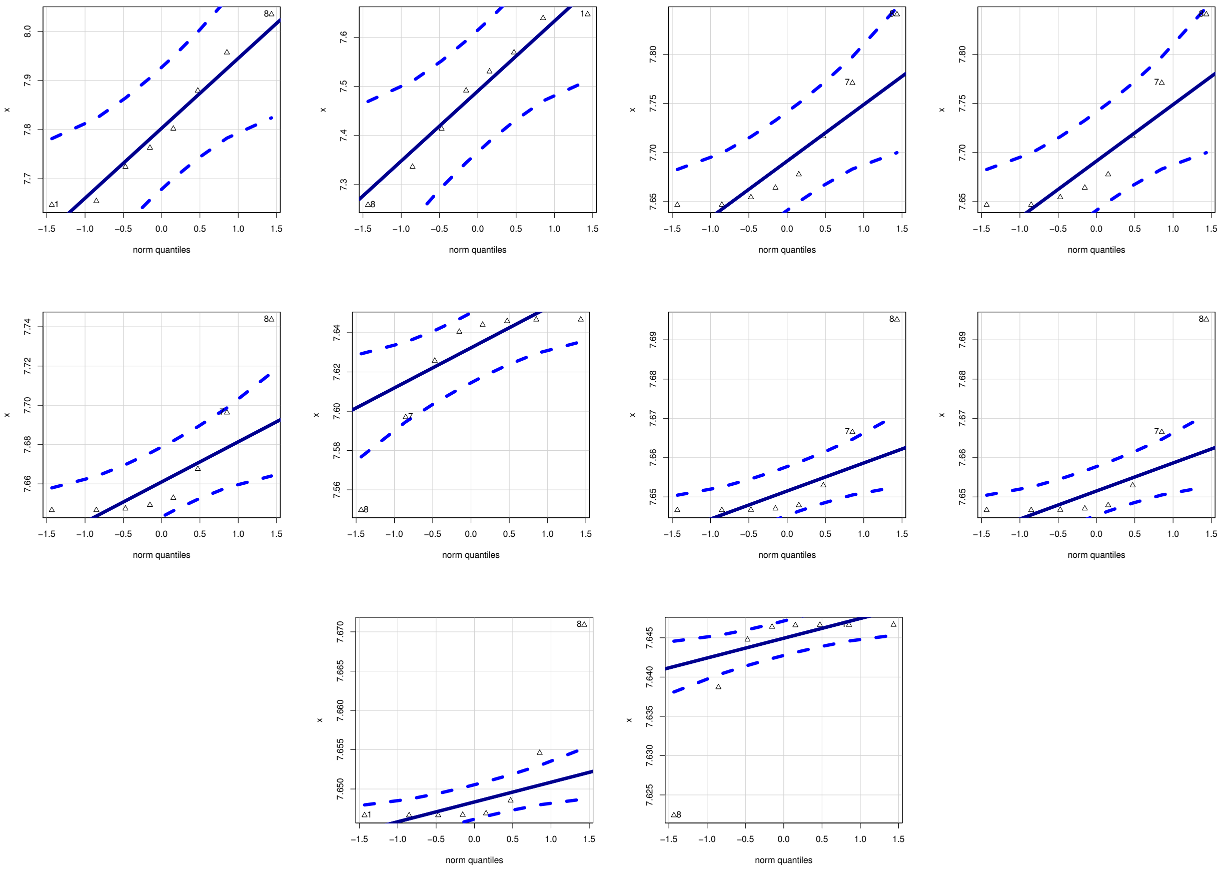

Figure 14. Q-Q analysis for the residuals.

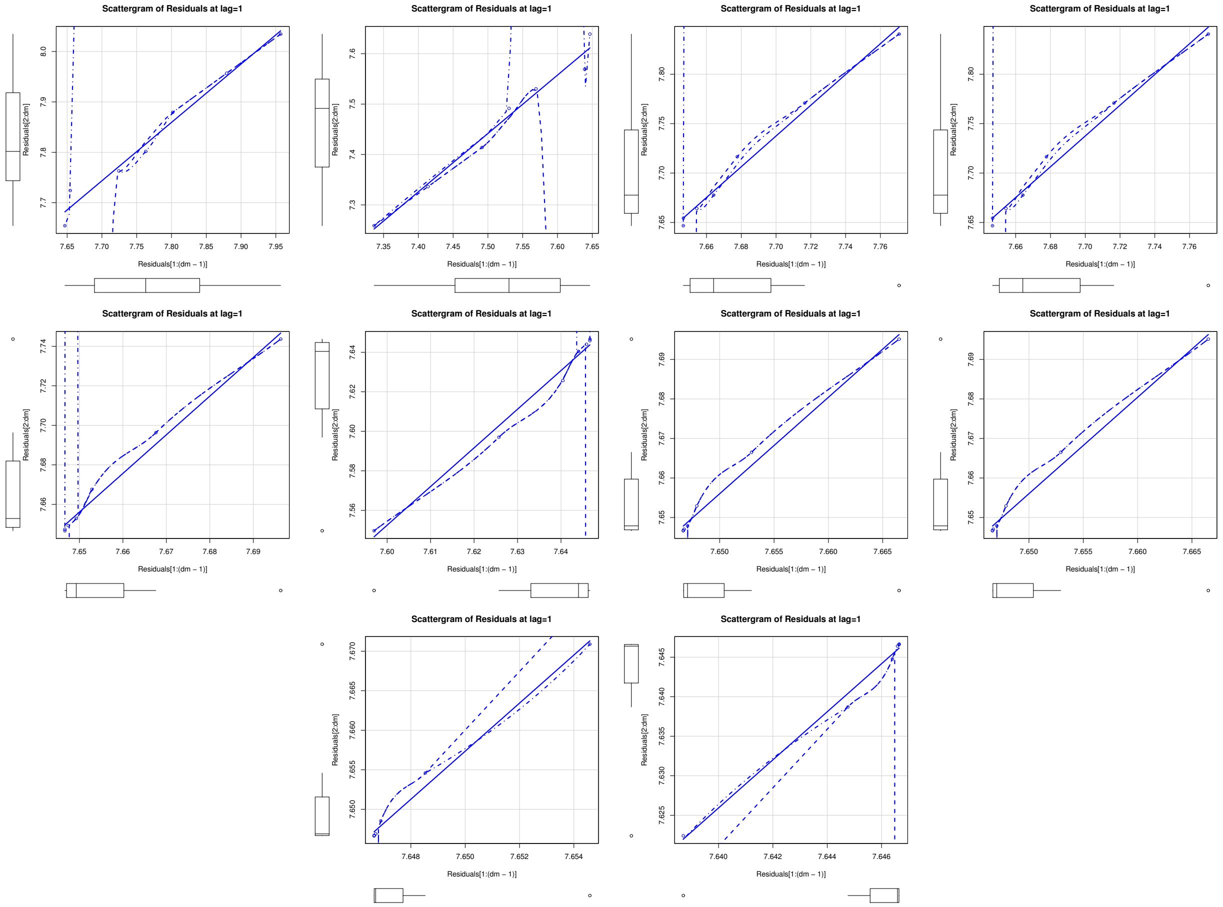

Figure 15. Scattergrams for the residuals.

Table 1: Point forecasting for the claim's payments in million.

	$r_{[k]}\|_{k\geq 1}$	$\xi_{[kk]}$	Point Forecasting				
$\vartheta\downarrow$	$r_{[1]},r_{[2]},r_{[3]},r_{[4]}$	$\xi_{[11]}$	$(Q_4)_{2013}$	$(Q_1)_{2014}$	$(Q_2)_{2014}$	$(Q_3)_{2014}$	$(Q_4)_{2014}$
10^{-4}	10^{-4},10^{-8},10^{-12},10^{-16}	10^{-4}	7.646723	7.646645	7.646645	7.646645	7.646645
10^{-2}	10^{-2},10^{-4},10^{-8},10^{-16}	10^{-2}	7.654403	7.646723	7.646646	7.646645	7.646645
0.10	0.1,10^{-2},10^{-3},10^{-4}	0.10	7.724225	7.654403	7.647421	7.646723	7.646653
0.15	$0.15, 0.023, 0.003, 0.001$	0.15	7.763015	7.664100	7.649263	7.647038	7.646704
0.20	$0.20, 0.04, 0.008, 0.002$	0.20	7.801805	7.677677	7.652851	7.647886	7.646893
0.30	$0.3, 0.09, 0.027, 0.008$	0.3	7.879384	7.716467	7.667592	7.652929	7.648530
0.40	$0.40, 0.16, 0.64, 0.026$	0.40	7.956964	7.770773	7.696296	7.666505	7.654589
0.50	$0.50, 0.063, 0.016, 0.004$	0.50	8.034544	7.840594	7.743620	7.695132	7.670889
-10^{-4}	-10^{-4},10^{-8},-10^{-12},10^{-16}	10^{-4}	7.646567	7.646645	7.646645	7.646645	7.646645
-10^{-2}	-10^{-2},10^{-4},-10^{-8},10^{-16}	10^{-2}	7.638887	7.646723	7.646644	7.646645	7.646645
-0.10	-0.1,10^{-2},-10^{-3},10^{-4}	0.10	7.569065	7.654403	7.645869	7.646723	7.646637
-0.15	$-0.15, 0.024, -0.003, 0.001$	0.15	7.530275	7.664100	7.644027	7.647038	7.646586
-0.20	$-0.20, 0.04, -0.008, 0.0016$	0.20	7.491485	7.677677	7.640439	7.647886	7.646397
-0.30	$-0.3, 0.09, -0.027, 0.0081$	0.3	7.413906	7.716467	7.625698	7.652929	7.64476
-0.40	$-0.40, 0.16-0.64, 0.026$	0.40	7.336326	7.770773	7.596994	7.666505	7.638701
-0.50	$-0.50, 0.06, -0.016, 0.004$	0.50	7.258746	7.840594	7.54967	7.695132	7.622401

Table 2: Forecasting residual analysis.

	$2013_{(Q_4)}$				$2014_{(Q_1)}$			
ϑ	$\mathrm{E}_{(Q_4)_{2013}}$	$\mathrm{APE}_{(Q_4)}$	$\|\mathrm{E}_{(Q_4)_{2013}}\|$	$\mathrm{E}^2_{(Q_4)_{2013}}$	$\mathrm{E}_{2014_{(Q_1)}}$	$\mathrm{APE}_{(Q_1)}$	$\|\mathrm{E}_{2014_{(Q_1)}}\|$	$\mathrm{E}^2_{2014_{(Q_1)}}$
10^{-4}	0.775720	0.09210154	0.775720	0.6017415	0.775798	0.09211080	0.775798	0.6018625
10^{-2}	0.768040	0.09118969	0.768040	0.5898854	0.775720	0.09210154	0.775720	0.6017415
0.10	0.698218	0.08289970	0.698218	0.4875084	0.768040	0.09118969	0.768040	0.5898854
0.15	0.659428	0.07829415	0.659428	0.4348453	0.758343	0.09003837	0.758343	0.5750841
0.20	0.620638	0.07368860	0.620638	0.3851915	0.744766	0.08842636	0.744766	0.5546764
0.30	0.543059	0.06447761	0.543059	0.2949131	0.705976	0.08382081	0.705976	0.4984021
0.40	0.465479	0.05526651	0.465479	0.2166707	0.651670	0.07737304	0.651670	0.4246738
0.50	0.387899	0.04605540	0.387899	0.1504656	0.581849	0.06908316	0.581849	0.3385483
-10^{-4}	0.775876	0.09212007	0.775876	0.6019836	0.775798	0.09211080	0.775798	0.6018625
-10^{-2}	0.783556	0.09303191	0.783556	0.6139600	0.775720	0.09210154	0.775720	0.6017415
-0.10	0.853378	0.10132191	0.853378	0.7282540	0.768040	0.09118969	0.768040	0.5898854
-0.15	0.892168	0.10592746	0.892168	0.7959637	0.758343	0.09003837	0.758343	0.5750841
-0.20	0.930958	0.11053301	0.930958	0.8666828	0.744766	0.08842636	0.744766	0.5546764
-0.30	1.008537	0.11974400	1.008537	1.0171469	0.705976	0.08382081	0.705976	0.4984021
-0.40	1.086117	0.12895510	1.086117	1.1796501	0.651670	0.07737304	0.651670	0.4246738
-0.50	1.163697	0.13816621	1.163697	1.3541907	0.581849	0.06908316	0.581849	0.3385483
$\sum$	12.41277	1.473773	12.41277	10.31905	11.52432	1.368288	11.52432	8.369748
Mean	0.775798	9.21108	0.775798	0.6449408	0.7202702	8.551797	0.7202702	0.5231093

4. Concluding remarks, future points and discussions

The future values of the expected claims are very important for the insurance companies for avoiding the big losses under uncertainty which may be produced from future claims. In this work, we defined a new size-of-loss synthetic autoregressive model ("SAR" for short) for the left skewed insurance claims datasets. The technique basically depends on exploring the time series insurance claims datasets under the all possible ARIMA models for selecting the best model. The SAR model is assessed due to some simulations experiments. The optimal parameter is also artificially determined. The insurance claims data is modeled using the synthetic autoregressive model. Many other

Table 2: Residuals analysis. Continued.

	$2014_{(Q_2)}$				$2014_{(Q_3)}$			
ϑ	$\text{E}_{(Q_2)_{2014}}$	$\text{APE}_{(Q_2)}$	$\lvert\text{E}_{(Q_2)_{2014}}\rvert$	$\text{E}^2_{(Q_2)_{2014}}$	$\text{E}_{(Q_3)_{2014}}$	$\text{APE}_{(Q_3)}$	$\lvert\text{E}_{(Q_3)_{2014}}\rvert$	$\text{E}^2_{(Q_3)_{2014}}$
10^{-4}	0.775798	0.0921108	0.775798	0.6018625	0.775798	0.0921108	0.775798	0.6018625
10^{-2}	0.775797	0.0921107	0.775797	0.6018610	0.775798	0.0921108	0.775798	0.6018625
0.10	0.775022	0.0920187	0.775022	0.6006591	0.775720	0.0921015	0.775720	0.6017415
0.15	0.773180	0.0917999	0.773180	0.5978073	0.775405	0.0920641	0.775405	0.6012529
0.20	0.769592	0.0913739	0.769592	0.5922718	0.774557	0.0919635	0.774557	0.5999385
0.30	0.754851	0.0896238	0.754851	0.5698000	0.769514	0.0913647	0.769514	0.5921518
0.40	0.726147	0.0862157	0.726147	0.5272895	0.755938	0.0897528	0.755938	0.5714423
0.50	0.678823	0.0805969	0.678823	0.4608007	0.727311	0.0863539	0.727311	0.5289813
-10^{-4}	0.775798	0.0921108	0.775798	0.6018625	0.775798	0.0921108	0.775798	0.6018625
-10^{-2}	0.775799	0.0921109	0.775799	0.6018641	0.775798	0.0921108	0.775798	0.6018625
-0.10	0.776574	0.0922029	0.776574	0.6030672	0.775720	0.0921015	0.775720	0.6017415
-0.15	0.778416	0.0924216	0.778416	0.6059315	0.775405	0.0920641	0.775405	0.6012529
-0.20	0.782004	0.0928477	0.782004	0.6115303	0.774557	0.0919635	0.774557	0.5999385
-0.30	0.796745	0.0945979	0.796745	0.6348026	0.769514	0.0913647	0.769514	0.5921518
-0.40	0.825449	0.0980059	0.825449	0.6813661	0.755938	0.0897528	0.755938	0.5714423
-0.50	0.872773	0.1036247	0.872773	0.7617327	0.727311	0.0863539	0.727311	0.5289813
$\sum$	12.41277	1.4737730	12.41277	9.654509	12.26008	1.455644	12.26008	9.398467
Mean	0.775798	9.211080	0.775798	0.6034068	0.766255	9.097778	0.766255	0.5874042

Table 2: Residuals analysis for $(Q_4)_{2014}$. Continued.

	$(Q_4)_{2014}$			
ϑ	$\text{E}_{(Q_4)_{2014}}$	$\text{APE}_{(Q_4)}$	$\lvert\text{E}_{(Q_4)_{2014}}\rvert$	$\text{E}^2_{(Q_4)_{2014}}$
10^{-4}	0.775798	0.09211080	0.775798	0.6018625
10^{-2}	0.775798	0.09211080	0.775798	0.6018625
0.10	0.775790	0.09210985	0.775790	0.6018501
0.15	0.775739	0.09210380	0.775739	0.6017710
0.20	0.775550	0.09208136	0.775550	0.6014778
0.30	0.773913	0.09188700	0.773913	0.5989413
0.40	0.767854	0.09116761	0.767854	0.5895998
0.50	0.751554	0.08923230	0.751554	0.5648334
-10^{-4}	0.775798	0.09211080	0.775798	0.6018625
-10^{-2}	0.775798	0.09211080	0.775798	0.6018625
-0.10	0.775806	0.09211175	0.775806	0.6018749
-0.15	0.775857	0.09211781	0.775857	0.6019541
-0.20	0.776046	0.09214025	0.776046	0.6022474
-0.30	0.777683	0.09233461	0.777683	0.6047908
-0.40	0.783742	0.09305400	0.783742	0.6142515
-0.50	0.800042	0.09498930	0.800042	0.6400672
$\sum$	12.41277	1.473773	12.41277	9.63111
Mean	0.775798	9.211080	0.775798	0.6019443

graphical techniques are considered such as the “nonparametric Kernel density estimation” for exploring initial insurance claims density shape, the “Quantile-Quantile” plot for exploring “normality” of the current data, the “total time on test” plot for exploring the initial shape of the empirical HRF, and the “box plot” for detecting the extreme claims. The SAR model is recommended for prediction under the left skewed insurance claims payment data. The main advantage of the new model depended on its simple procedures in prediction especially if it has only one parameter. We hope that the SAR model will attract more researchers in future works. A potential study on value-at-risk estimation based on the PORT mean-of-order-p methodology is considered.

Finally, here are the highly-rated recommendations tailored specifically to help motor insurance companies improve their operations, forecasting accuracy, risk management, and reserving practices , particularly in the U.K. Motor Non-Comprehensive line of business. Implement SAR(p) models for accurate short-term forecasting of motor insurance claims payments, especially when dealing with left-skewed historical data. Leverage ARIMA-based approaches to model time-dependent claim development patterns and enhance claims reserving accuracy over multiple quarters. Utilize graphical tools such as Cullen and Frey plots , Q-Q plots , Kernel density estimation , and box plots to understand the distributional properties of claims data before modeling. Perform thorough residual analysis (e.g., ACF/PACF plots, scattergrams) to ensure that the selected SAR or ARIMA model adequately captures temporal dependencies in the data. Given that claims data is often reported quarterly, build separate time series models for each quarter to better capture seasonality and trend components . Apply point forecasting techniques along with error metrics like MAPE , MSE , and APE to quantify forecast accuracy and support financial planning. Use predicted claims values to estimate solvency capital requirements and maintain adequate reserves to meet future obligations under uncertainty. Incorporate historical claim trends identified through time series analysis into premium pricing strategies to reflect actual risk exposure more accurately. Move beyond traditional deterministic methods like chain-ladder; automate reserving using statistical models such as SAR and ARIMA to reduce human bias and increase precision. Monitor deviations from expected claim development patterns using time series diagnostics to detect fraudulent activities , operational inefficiencies, or underwriting anomalies. Create dashboards integrating time series forecasts, residuals, and diagnostic plots to support real-time monitoring and decision-making in claims departments. Invest in training actuaries and risk analysts in modern statistical forecasting methods, including autoregressive modeling , to keep pace with evolving actuarial science. These recommendations are designed to help U.K. motor insurers , particularly those dealing with Non-Comprehensive policies , make data-driven decisions in areas such as reserving, pricing, solvency, fraud detection, and operational efficiency . By embracing advanced time series modeling techniques like SAR and ARIMA, insurers can significantly enhance their ability to manage uncertainty, comply with regulations, and optimize financial performance.

This work is highly significant in the field of insurance as it provides an effective tool for predicting future claim liabilities, which is essential for accurate reserving and risk management. The proposed SAR model allows insurers to analyze historical claims data and generate reliable forecasts, helping them avoid financial losses due to uncertainty. By applying time series techniques such as ARIMA and SAR modeling, the study enhances traditional actuarial methods used in claims forecasting. The paper uses real-world U.K. Motor Non-Comprehensive insurance claims data from 2007 to 2013, making it directly applicable to industry practices. Forecasting future payments enables insurers to maintain solvency by setting aside adequate reserves and optimizing capital allocation. The SAR(1) model, in particular, demonstrates strong forecasting accuracy, especially for short-term predictions like quarterly claims. These forecasts are crucial for meeting regulatory requirements such as Solvency II, which mandates precise estimation of outstanding liabilities. The paper contributes to the advancement of predictive analytics in insurance by combining classical actuarial science with modern statistical modeling. It also encourages further research into parametric time series models tailored for skewed insurance data. The ability to simulate artificial data and estimate parameters artificially supports broader applications across different lines of insurance. The methodology can be extended to other types of insurance such as health, life, and property, where similar forecasting challenges exist. Moreover, the integration of statistical forecasting into pricing strategies improves premium rate-setting based on actual claim trends. The study emphasizes the importance of identifying seasonality, trend, and autocorrelation in claims data to improve forecasting precision. Ultimately, adopting such advanced modeling techniques empowers insurers to make data-driven decisions, enhance operational efficiency, and strengthen financial planning. The paper sets a foundation for future innovation in actuarial science and promotes collaboration between academia and the insurance industry.

It is important to acknowledge that the insurance data used in this study may be somewhat dated. This is primarily due to the inherent difficulty in accessing recent, high-quality insurance datasets, as such data is often confidential, proprietary, or not readily available for public research. The dataset employed in this analysis serves as a practical illustrative example, intended to demonstrate the application of actuarial and statistical methods in a real-world insurance context. While the data may not reflect the most current market conditions, it remains valuable for exploring the dynamics of claim development and testing the proposed modeling approach. We hope that future papers will be conducted using more recent and extensive insurance claims data, which would enhance the relevance and applicability of the findings. The primary insurance-related objective of this research is to introduce a novel predictive framework that could assist insurance companies in improving their financial forecasting of claim liabilities. By offering an innovative modeling perspective, we aim to contribute to the ongoing efforts in the industry to enhance reserve accuracy, strengthen solvency planning, and support data-driven decision-making processes.

REFERENCES

1. Aboraya, M., Ali, M. M., Yousof, H. M. and Ibrahim, M. (2022). A New Flexible Probability Model: Theory, Estimation and Modeling Bimodal Left Skewed Data. Pakistan Journal of Statistics and Operation Research, 18(2), 437-463.
2. Aboraya, M., Ali, M. M., Yousof, H. M. and Ibrahim, M. (2022). A Novel Lomax Extension with Statistical Properties, Copulas, Different Estimation Methods and Applications. Bulletin of the Malaysian Mathematical Sciences Society, (2022) https://doi.org/10.1007/s40840-022-01250-y
3. Al-babtain, A. A., Elbatal, I. and Yousof, H. M. (2020). A New Flexible Three-Parameter Model: Properties, Clayton Copula, and Modeling Real Data. Symmetry, 12(3), 440.
4. Al-Babtain, A. A., Elbatal, I. and Yousof, H. M. (2020). A new three parameter Fréchet model with mathematical properties and applications. Journal of Taibah University for Science, 14(1), 265-278.
5. Box, G. and Jenkins, G. (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.
6. Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
7. Charpentier, A. (2014). Computational actuarial science with R. CRC press.
8. Cummins, J. D. and Griepentrog, G. L. (1985). Forecasting automobile insurance paid claim costs using econometric and ARIMA models. International Journal of Forecasting, 1(3), 203-215.
9. Darekar, A. and Reddy, A. (2017). Forecasting oilseeds prices in India: Case of groundnut. Forecasting Oilseeds Prices in India: Case of Groundnut (December 14, 2017). J. Oilseeds Res, 34(4), 235-240.
10. Elgohari, H. and Yousof, H. M. (2020). A Generalization of Lomax Distribution with Properties, Copula and Real Data Applications. Pakistan Journal of Statistics and Operation Research, 16(4), 697-711.
11. Elgohari, H. and Yousof, H. M. (2020). New Extension of Weibull Distribution: Copula, Mathematical Properties and Data Modeling. Statistics, Optimization & Information Computing, 8(4), 972-993.
12. El-Morshedy, M., Alshammari, F. S., Hamed, Y. S., Eliwa, M. S., Yousof, H. M. (2021). A New Family of Continuous Probability Distributions. Entropy, 23, 194.
13. Goual, H., Yousof, H. M., & Ali, M. M. (2019). Validation of the odd Lindley exponentiated exponential by a modified goodness of fit test with applications to censored and complete data. Pakistan Journal of Statistics and Operation Research, 15(3), 745-771.
14. Hafiz, U. A., Salleh, F., Garba, M. and Rashid, N. (2021). Projecting insurance penetration rate in Nigeria: An ARIMA approach. REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS, 11(3), 63-75.
15. Hamedani, G. G., Altun, E., Korkmaz, M. Ç., Yousof, H. M., & Butt, N. S. (2018). A new extended G family of continuous distributions with mathematical properties, characterizations and regression modeling. Pakistan Journal of Statistics and Operation Research, 737-758.
16. Hamedani, G. G., Korkmaz, M. C., Butt, N. S. and Yousof, H. M. (2021). The Type I Quasi Lambert Family: Properties, Characterizations and Different Estimation Methods. Pakistan Journal of Statistics and Operation Research, 17(3), 545-558.
17. Hamedani, G. G., Rasekhi, M., Najibi, S., Yousof, H. M., & Alizadeh, M. (2019). Type II general exponential class of distributions. Pakistan Journal of Statistics and Operation Research, 15(2), 503-523.
18. Hamedani, G. G., Yousof, H. M., Rasekhi, M., Alizadeh, M. and Najibi, S. M., (2018). Type I general exponential class of distributions. Pakistan Journal of Statistics and Operation Research, 14(1), 39-55.
19. Ibrahim, M., Aidi, K., Ali, M. M. and Yousof, H. M. (2022). A Novel Test Statistic for Right Censored Validity under a new Chen extension with Applications in Reliability and Medicine. Annals of Data Science, forthcoming. doi.org/10.1007/s40745-022-00416-6
20. Ibrahim, M., Yadav, A. S., Yousof, H. M., Goual, H., & Hamedani, G. G. (2019). A new extension of Lindley distribution: modified validation test, characterizations and different methods of estimation. Communications for Statistical Applications and Methods, 26(5), 473-495.
21. Iqbal, M. C., Jamshaid, T. M., & Rashid, A. Q. A. (2016). Forecasting of wheat production: a comparative study of Pakistan and India. International Journal of Advanced Research (IJAR), 4(12), 698-709.
22. Jakaša, T., Andročec, I. and Sprčić, P. (2011). Electricity price forecasting-ARIMA model approach. In 2011 8th International Conference on the European Energy Market (EEM) (pp. 222-225). IEEE.
23. Jang, K. P., Kam, S. and Park, J. Y. (1991). Trend and Forecast of the Medical Care Utilization Rate, the Medical Expense per Case and the Treatment Days per Cage in Medical Insurance Program for Employees by ARIMA Model. Journal of Preventive Medicine and Public Health, 24(3), 441-458.
24. Korkmaz, M. Ç., Altun, E., Chesneau, C. and Yousof, H. M. (2022). On the unit-Chen distribution with associated quantile regression and applications. Mathematica Slovaca, 72 (2022), No. 3, 765-786.

25. Kumar, V. S., Satpathi, D. K., Kumar, P. P. and Haragopal, V. V. (2020). Forecasting motor insurance claim amount using ARIMA model. In AIP Conference Proceedings (Vol. 2246, No. 1, p. 020005). AIP Publishing LLC.
26. Mansour, M. M., Ibrahim, M., Aidi, K., Butt, N. S., Ali, M. M., Yousof, H. M., & Hamed, M. S. (2020). A New Log-Logistic Lifetime Model with Mathematical Properties, Copula, Modified Goodness-of-Fit Test for Validation and Real Data Modeling. Mathematics, 8(9), 1508.
27. Mansour, M., Rasekhi, M., Ibrahim, M., Aidi, K., Yousof, H. M., & Elrazik, E. A. (2020). A New Parametric Life Distribution with Modified Bagdonavičius–Nikulin Goodness-of-Fit Test for Censored Validation, Properties, Applications, and Different Estimation Methods. Entropy, 22(5), 592.
28. Mansour, M. M., Butt, N. S., Yousof, H. M., Ansari, S. I., & Ibrahim, M. (2020). A Generalization of Reciprocal Exponential Model: Clayton Copula, Statistical Properties and Modeling Skewed and Symmetric Real Data Sets. Pakistan Journal of Statistics and Operation Research, 16(2), 373-386.
29. Mohammadi, H. and Rich, D. P. (2013). Dynamics of unemployment insurance claims: an application of ARIMA-GARCH models. Atlantic Economic Journal, 41(4), 413-425.
30. Nascimento, A. D., Silva, K. F., Cordeiro, G. M., Alizadeh, M., Yousof, H. M., & Hamedani, G. G. (2019). The odd Nadarajah-Haghighi family of distributions: properties and applications. Studia Scientiarum Mathematicarum Hungarica, 56(2), 185-210.
31. Nath, B., Dhakre, D. S. and Bhattacharya, D. (2019). Forecasting wheat production in India: An ARIMA modelling approach. Journal of Pharmacognosy and Phytochemistry, 8(1), 2158-2165.
32. Palakuru, M., Yarrakula, K., Chaube, N. R., Sk, K. B. and Satyaji Rao, Y. R. (2019). Identification of paddy crop phenological parameters using dual polarized SCATSAT-1 (ISRO, India) scatterometer data. Environmental Science and Pollution Research, 26(2), 1565-1575.
33. Sahu, P. K., Mishra, P., Dhekale, B. S., Vishwajith, K. P. and Padmanaban, K. (2015). Modelling and forecasting of area, production, yield and total seeds of rice and wheat in SAARC countries and the world towards food security. American Journal of Applied Mathematics and Statistics, 3(1), 34-48.
34. Shehata, W. A. M. and Yousof, H. M. (2021). The four-parameter exponentiated Weibull model with Copula, properties and real data modeling. Pakistan Journal of Statistics and Operation Research, 17(3), 649-667.
35. Shehata, W. A. M., Butt, N. S., Yousof, H., & Aboraya, M. (2022). A New Lifetime Parametric Model for the Survival and Relief Times with Copulas and Properties. Pakistan Journal of Statistics and Operation Research, 18(1), 249-272.
36. Shrahili, M.; Elbatal, I. and Yousof, H. M. Asymmetric Density for Risk Claim-Size Data: Prediction and Bimodal Data Applications. Symmetry 2021, 13, 2357.
37. Venezian, E. C. and Leng, C. C. (2006). Application of spectral and ARIMA analysis to combined-ratio patterns. The Journal of Risk Finance.
38. Yadav, A. S., Altun, E., & Yousof, H. M. (2021). Burr–Hatke Exponential Distribution: A Decreasing Failure Rate Model, Statistical Inference and Applications. Annals of Data Science, 8(2), 241–260.
39. Yousof, H. M., Ali, M. M., Hamedani, G. G., Aidi, K. & Ibrahim, M. (2022). A new lifetime distribution with properties, characterizations, validation testing, different estimation methods. Statistics, Optimization & Information Computing, 10(2), 519-547.
40. Yousof, H. M., Korkmaz, M. Ç., K., Hamedani, G. G and Ibrahim, M. (2022). A novel Chen extension: theory, characterizations and different estimation methods. Eur. J. Stat, 2(2022), 1-20.