Tanzania Journal of Science 48(1): 13-21, 2022
ISSN 0856-1761, e-ISSN 2507-7961

Distribution and Pattern of an Insurance Health Claim System: A Time Series Approach

Maxwell Mashasha[1], Praise Mutize[1] and Felix Mazunga[2]
[1] *Department of Applied Mathematics and Statistics, Midlands State University, Zimbabwe*
[2]*Department of Applied Physics and Telecommunications, Midlands State University, Zimbabwe*
Corresponding author: mashashamn@staff.msu.ac.zw
Received 11 Nov 2021, Revised 8 Mar 2022, Accepted 9 Mar 2022, Published Mar 2022
DOI: https://dx.doi.org/10.4314/tjs.v48i1.2

Abstract

There is a continuous increase in health costs, thereby increasing pressure on individuals and consequently making the amounts claimed by the insured to be on the increase. In this study, data was collected from a large local insurance company in Zimbabwe for the period from January 2012 to December 2016. The aim of this study was to analyse the distribution and future pattern of insurance health claim system using time series approach. Akaike information criterion and Schwarz Bayesian criterion were used to select the adequate model through maximum likelihood estimation methods. ARIMA (0, 0, 0) (1, 0, 1) [12] is the model that was chosen to forecast claim amounts. The use of ARIMA models proves to be an excellent instrument for predicting and capturing the cost trend of health claims which can help in decision making to insurance companies.

Keywords: ARIMA; Box-Jenkins; health insurance; time series.

Introduction

In the field of actuarial science, general insurance is the most growing area. General insurance includes personal, health, property, commercial risks and liability insurance, among others. Insurance policies are used to hedge against financial loss risks for the insured goods or people. Health insurance is an acknowledged type of insurance coverage that protects insured persons from paying high treatment costs in the event of sickness (Bietsch et al. 2020). Insurance can be delivered through private and public markets and the most common types of private insurance compensation are hospital costs (medical expenses).

There is a common worldwide problem of high costs of health care services, caused by the increasing demands for these services (Jiying et al. 2019). Health insurance bills have been exponentially increasing in recent years, with additional costs being passed down to the consumers as shortfalls (Bietsch et al. 2020). The linear pricing model is commonly used in the pricing of products offered. Mwangi and Murigu (2015) highlighted many factors that affect insurance claims and these include claim size/amount, size of the insurance company, type of management, leverage and liquidity among others. Claim settlement is the monetary compensation done by the insurer to the insured for the unfortunate event, it is mainly affected by the insurer's liquidity and how convenient the insured is in paying the premiums as agreed in the policy (Mwangi and Murigu 2015).

The National Health Strategy (NHS) was introduced in 2003, accounting for a significant increase in health services in Zimbabwe. The strategy revealed that during the hyperinflationary period, those in the

political circles and the financially stable individuals have been flying to different nations for health services. During the hyperinflationary period, the economy was not certain; hence insurance business in Zimbabwe was fluctuating. The insurance sector was revived when the multicurrency system was adopted in Zimbabwe. Insurance data contain irregular claims with large sums of money; hence Boland (2007) highlighted the need to find suitable skewed statistical distributions that will be used to make suitable decisions on premium loadings, expected profits and reserves with greater profitability. In order to predict the future of the insurance sector some statistical approaches which were used include time series analysis and econometric modelling (Cummins 1973, Jiying et al. 2019). In this study, we use a time series approach to predict the pattern and distribution of health claim system based on a Zimbabwean health insurance company.

The recent rise in the costs of health care has become a major challenge incapacitating individuals and insurance organizations as they have limited models for claim prediction. To find the appropriate distribution, there is need for identification of significant factors that affect the insurance health claims. Time Series methods are used to explain, clarify, forecast, and control changes through time for selected variables (Jiying et al. 2019). In the world of spectrum analysis, time series is divided into separate components by breaking down series into some wave fronts and combinations of basic Fourier series (Lancaster et al. 2018). The analysis involved use of power spectral, thus plotting variance in the series versus frequency that can be applied in autoregressive time series. Several time series methods have been established with the concept of autoregressive (AR) and moving average (MA) models being formulated by decomposing ARIMA (p, d, q) into AR and MA. The AR captures the association between the present values of the time series and its previous values where AR (1) shows that present observation is correlated at time with its historical values. The MA component signifies the duration of the effect of a random shock, while the autocorrelation (ACF) and partial autocorrelation (PACF) plots are used to estimate the values of p and q. Box-Jenkins approach to modelling ARIMA (p, d, q) process is used in this research.

Chan (2004) used at least two time-series modelling techniques proposed by Tiao (1985), to construct a stochastic funding model for price increase, share dividends, percentage dividend yields, and long-term interest costs inside the United Kingdom. Pflaumer (1992) showed that Box-Jenkins approach is comparable to a simple trend model that makes long range predictions when he managed to forecast the population of the United States up to the year 2080 using the Box Jenkins approach. Literature had proved that the Box-Jenkins method is more dependable than the traditional demographic methods.

ARIMA method for constructing time series models was proposed by the two mathematicians Box and Jenkins (1976). These mathematicians in 1970 had earlier developed the Box Jenkins model in the publication titled "Time Series Analysis: Forecasting and Control". Guiahi (2000) looked on issues and methodologies for fitting samples of insurance data with alternative statistical distributions in greater detail. Meyers (2005) used actuarial modelling approach to fit a statistical distribution to 250 claims. Log-normal, gamma and weibull distributions were evaluated in this study using the maximum likelihood estimation method. Lee and Miller (2002) used stochastic time series models to project the upward thrust in health expenditure in the US and they found out that the healthcare spending would rise by around 8% of GDP from the current 2.2% by 2075. El-Bassiouni and El-Habashi (1991) forecast monthly obligatory motor insurance claims using Box Jenkins methodology, demonstrating that their model can accurately predict annual potential claims. Using ARIMA models, Zheng et al. (2020) estimated China's total health spending and ARIMA (5, 1) model proved to be most

suitable to forecast total expenditure as a percentage of GDP.

Renshaw (2004) expanded quasi-likelihood in claim amounts for non-life insurance and used 96 maximum likelihood estimates to fit the best model since quasi-likelihood parameter 97 estimates have comparable asymptotic properties to the maximum likelihood parameter 98 estimates. In this study, we used the maximum likelihood approach to estimate model parameters, and also the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) to get the best probability distribution that matches the data set. Amounts claimed from a health insurance institution were analysed using time series methods. The use of seasonal ARIMA models proves to be an excellent instrument for predicting and capturing the cost trend of health claims which can help in decision making to insurance companies.

Materials and Methods

This is a retrospective cohort study, where we used a quantitative research approach to accomplish the main goal of the study. Secondary data from 2012 to 2016 were used to determine future patterns and distribution of claims. The variables of interest are: age, gender, medical aid scheme (society) type, date of claim and amount claimed. In this research, a time series modelling approach was used to fit a distribution to 3664 claims from a Zimbabwean local insurance company and all the data cleaning procedures which include removal of unwanted observations, fixing outliers and handling missing values were done. The Box Jenkins approach was used, which involved model identification where we convert non stationary data into stationary series. This was done using the unit root test hypothesis and the decision is to "reject if $p > \alpha = 0.05$". Other models that were used are the autoregressive process (AR), moving averages (MA), and autoregressive integrated moving average (ARIMA). ACF and PACF plots were used to test for autocorrelation since they give a pictorial view of autocorrelation patterns in a series trend. Akaike Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC) were used to calculate number of autoregressive and moving average terms and a good model has the lowest value.

Model diagnostics involves evaluating the significance of the coefficients, order of the model, and also estimating residual behaviour. If the calculated model satisfies adequately the distribution from which the data come from, the residuals, should be independent of one another, mean and variance must be constant over time (the white noise process where $Z_t \sim N(0, \sigma^2)$). Model adequacy can be tested using the Ljung Box test. Evaluation of normality of data is a precondition for many statistical tests as normal data is a basic assumption in statistics. Several methods exist to test data normality, the most commonly used are Shapiro–Wilk test, Kolmogorov–Smirnov test, skewness and kurtosis. Histograms, box plots, P–P plots, and Q–Q plots are diagnostic methods used. In this research, we utilize the Kolmogorov-Smirnov test since our sample size is big. For model diagnostics, histograms, the Jarque Bera normality test, box plots, Q-Q plots are commonly used. In all analyses we mainly used the R package.

Results and Discussion

Descriptive statistics

The plot in Figure 1 shows a clear fluctuating pattern of insurance claims from 2012 to 2016. The fluctuations seem to be seasonal. There are approximately two peaks per annum; hence these data are ideal for time series analysis because of the presence of a time trend.

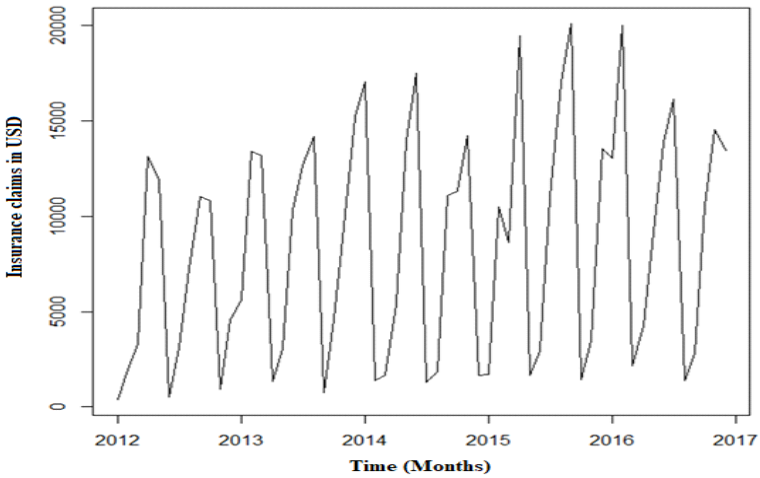

Figure 1: Plot of insurance claims against time.

Monthly distribution of claims amounts

Figure 2 (box plots) summarises the claims data in a box and whisker plot for identifications of patterns in the insurance industry. It shows the monthly claims distribution and claim skewness. The minimum claim amounts are much less than $5000, as can be seen in the 1st and 6th and maximum claims of up to $20000 were recorded in the 2nd and 9th month of the year. In other months there are long upper whiskers, suggesting claimed amounts are skewed to the upper side.

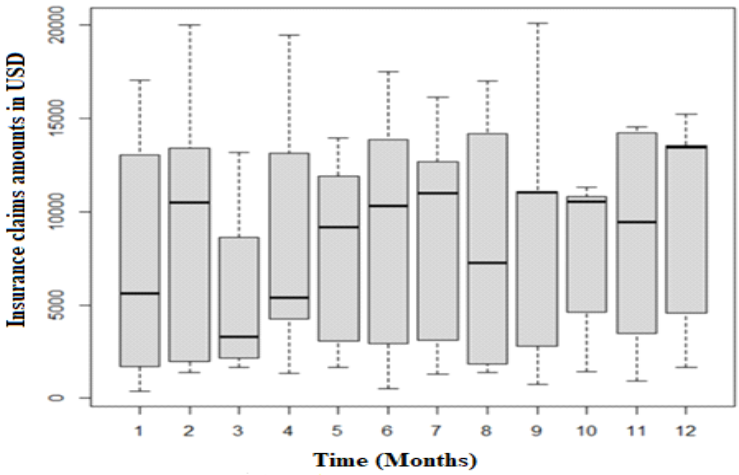

Figure 2: Box and Whisker representation of claims amounts.

Correlograms of insurance claims

In Figure 3, the ACF suggests an MA (3), since there are three significant spikes (number of spikes above base line) that are out of the confidence interval at sequential lags 0.4, 0.8 and 1.2 and the PACF plot suggests an AR (1) because there is only one spike displaying a positive autocorrelation depicting stationarity.

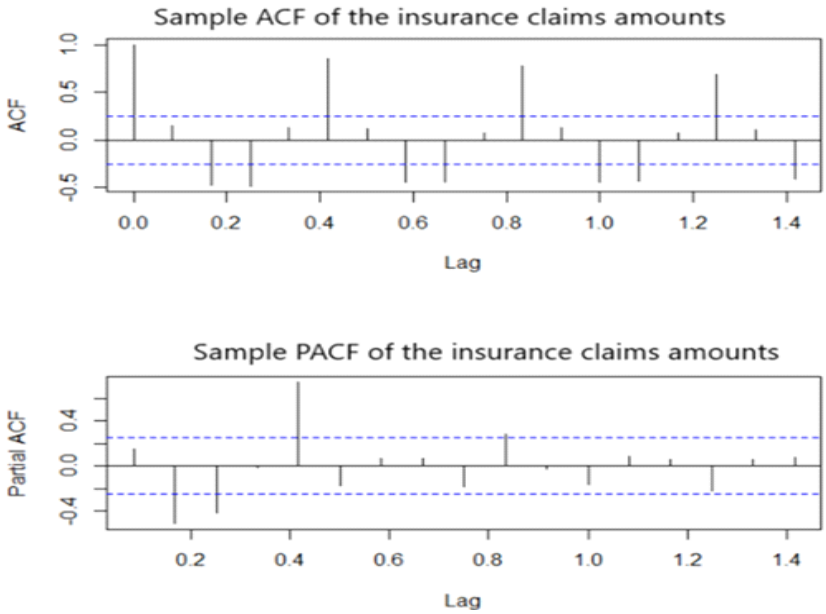

Figure 3: Correlograms.

Test for stationarity

Augmented Dickey-Fuller Test

H_0: The data is a non-stationary series

H_1: The data is a stationary series

Table 1: Unit root test for checking stationarity

Dickey fuller	Lag order	p-value
-10.257	3	0.01

The ADF test in Table 1 suggests that the time series of claims amounts is really significant at lag 3 which shows that the data is stationary; the p-value is less than 0.05.

Conclusion: we reject H_0 and conclude that the data is a stationary series at lag 3 and there is no unit root since $p = 0.01 < 0.05$ and it is significant.

Test for normality

Since we have a bigger sample size (> 50), we considered the Kolmogorov–Smirnov test for checking the normality assumption.

H_0 : Data is from a normally distributed population

H_1: Data is not from a normally distributed population

Decision criterion: If $p > 0.05$, we fail to reject H_0 and conclude that data is from a normally distributed population.

Conclusion: We fail to reject H_0 since $p = 0.1022 > 0.05$ and conclude that the data is from a normally distributed population.

The normal plot shows the residuals against their expected values, the normal probability plot of residuals approximately follows a pattern as shown in Figure 4(a). Histograms (Figure 4(b)) depicts how well models match the data but, in our scenario, they are skewed to the left. The normal Q-Q plot (Figure 4(c)) indicates that the process is normally distributed since most of the quantiles lie roughly on the 45-degree reference line (straight line) while the box plot (Figure 4(d) shows that the median (interquartile range) amount is represented by the bold black line on the box plot which is slightly below \$10000. The estimates of the 1st and 3rd quartiles are \$2000 and \$14000, respectively.

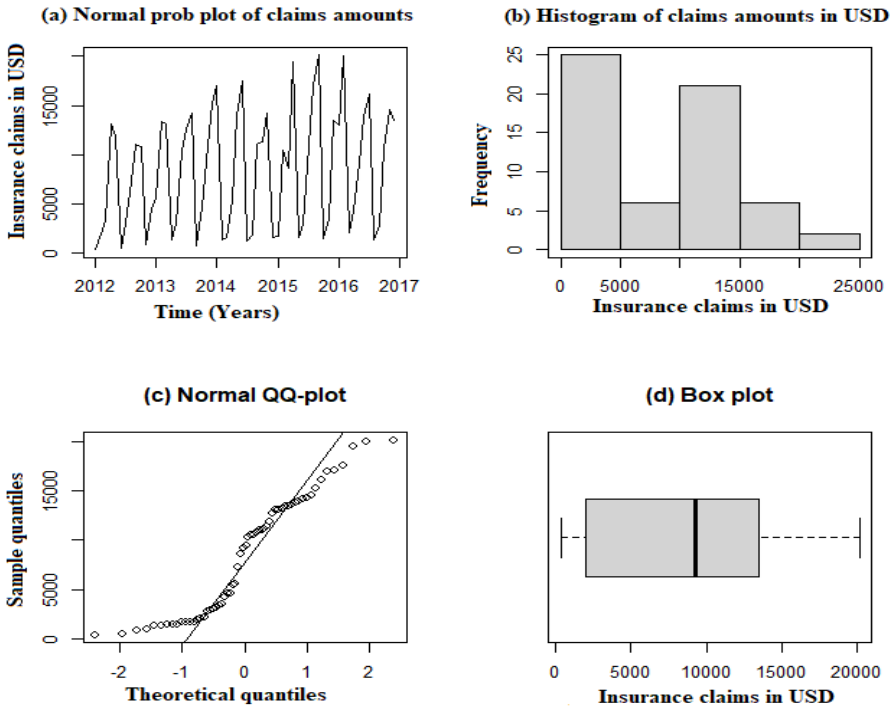

Figure 4: Residual analysis.

Model selection using the maximum likelihood method

The best model to be used for future forecasting has the minimum AIC and BIC values. In this case, ARIMA (0, 0, 0) (1, 0, 1) [12] model is the one with the smallest AIC of 1199.21, also corresponding to the minimum BIC of 1207.587 as shown in Table 2.

Table 2: Akaike and Bayesian information criterion for ARIMA model selection

MODEL	AIC	BIC
ARIMA (2,0,2)	1255.872	1268.438
ARIMA (0,0,0)	1217.88	1222.069
ARIMA (1,0,0)	1218.5	1224.783
ARIMA (0,0,1)	1215.35	1221.633
ARIMA (1,0,1)	1212.502	1220.88
ARIMA (2,0,2) (1,0,1) [12]	Inf	Inf
ARIMA (1,0,0) (1,0,0) [12]	1205.045	1213.423
ARIMA (0,0,1) (0,0,1) [12]	Inf	Inf
ARIMA (1,0,0) (1,0,1) [12]	1201.184	1211.656
ARIMA (1,0,0) (0,0,1) [12]	Inf	Inf
ARIMA (0,0,0) (1,0,1) [12]	1199.21	1207.587
ARIMA (0,0,0) (0,0,1) [12]	Inf	Inf
ARIMA (0,0,0) (1,0,0) [12]	1203.073	1209.356
ARIMA (0,0,1) (1,0,1) [12]	1201.163	1211.635
ARIMA (1,0,1) (1,0,1) [12]	1202.238	1214.804
ARIMA (0,0,0) (1,0,1) [12]	Inf	Inf

Model diagnostic

The standardized residuals (Figure 5 (top)) show a model of ARIMA (0,0,0) (1,0,1) [12] fitted to the amounts of insurance claims. Because of the lack of trend, the plot supports the model and it shows no apparent pattern in claims. The plot of ACF residuals against lag plot (Figure 5 (middle)) shows the residuals from ARIMA (0,0,0) (1,0,1) [12] model of health insurance claims amounts, concluding that the plot shows a substantial statistically relevant witness of non-zero residual autocorrelation. Autocorrelations for the residual's series are non-significant. In the plot of the Ljung-Box statistics (Figure 5 (bottom)), it is observed that some of the residuals lie above blue line and some below. Thus, the autocorrelations of the time series are different from zero. The residuals are random indicating that the model provides an adequate fit to the data series.

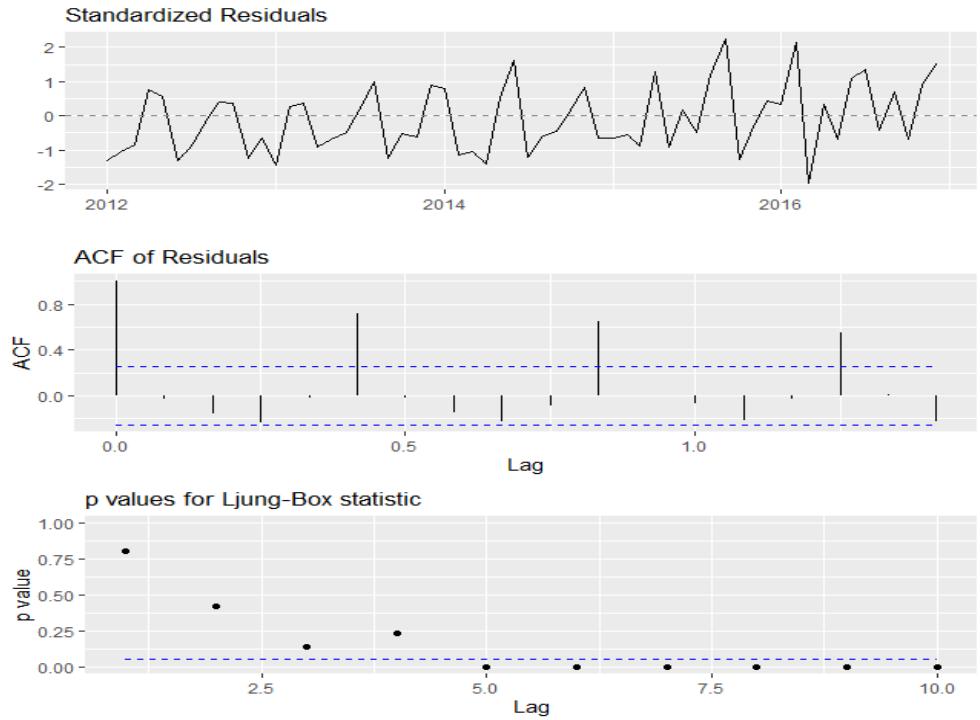

Figure 5: Plots of standardized residuals, ACF residuals against lag and Ljung-Box statistics.

Forecasting

The dark area in Figure 6 shows the predicted health insurance claims amounts using the ARIMA (0, 0, 0) (1, 0, 1) [12] model. The years 2012 to 2016 represent the time component which is important in predicting claims amounts in the coming years. An increase in claims was expected due to the national economic instability, which will later fluctuate and stabilize after the year 2018. Forecasting broaden recursion forwards of estimation of claims amounts, and allows projections, improvements and decision making in the insurance industry.

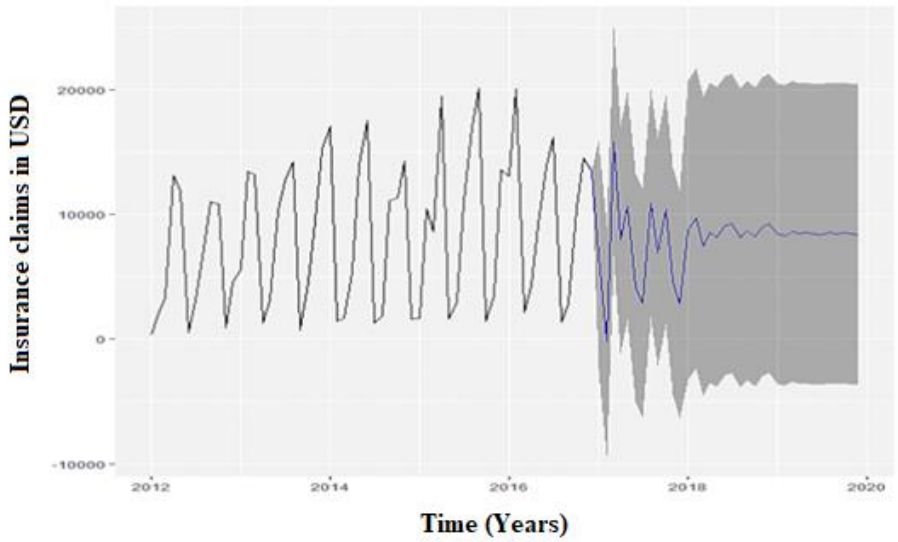

Figure 6: Forecasting health insurance claims amounts.

Conclusions

This research was done to analyse the distribution and future pattern of insurance health claim system using a time series approach. We used secondary data from 2012 to 2016 from a Zimbabwean local insurance company to determine future pattern and distribution of claims. The Kolmogorov–Smirnov test was adopted for testing the normality assumption. ARIMA model was developed for predicting time series plots. Normal probability plot, Q-Q plot, histogram and box plot were used for residual analysis. We found that claims amounts were normally distributed, but uncorrelated over time with noticeable skewness. The AIC and BIC were used to select the model that adequately fits the data using the maximum likelihood estimation method from Box-Jenkins methodology.

Acknowledgements: The authors would like to thank Midlands State University for promoting this study.

Declaration: The authors declare that there is no conflict of interest whatsoever regarding this research work. No funding has been received for this work.

References

Bietsch K, Rosenberg R, Stover J and Winfrey W 2020 Determinants of health insurance coverage and out-of-pocket payments for health care in Jordan: secondary analysis of the 2017-18 JPFHS. *DHS Further Analysis Reports No. 138*. Rockville, Maryland, USA ICF. Available at https://www.dhsprogram.com/pubs/pdf/FA138/FA138.pdf. Accessed 24th August 2021.

Boland PJ 2007 Statistical methods in insurance and actuarial science. CRC Press.

Box GE and Jenkins GM 1976 Time series analysis, control, and forecasting. San Francisco, CA: Holden Day.

Chan NH 2004 Time series: applications to finance. John Wiley & Sons.

Cummins JD 1973 An econometric model of the life insurance sector of the U.S. economy. *J. Risk and Insur.* 40(4): 533-554.

El-Bassiouni MY and El-Habashi MH 1991 Forecasting compulsory motor insurance claims in Kuwait. *Insur Math Econ.* 10(2): 85-92.

Guiahi F 2000 Fitting to loss distributions with emphasis on rating variables. Available at

www.casact.org/pubs/forum/01wforum/01wf133.pdf. Accessed 20th August 2021.

Jiying W, Beraud JD, Mensah IA 2019 Managing and predicting the number of health insurance claims in Ghana based on big data and time series analysis: a case study of Kumasi Metropolis, Ghana. *Int. J. Bus. Manag.* 7(3): 97-107.

Lancaster G, Iatsenko D, Pidde A, Ticcinelli V and Stefanovska A 2018 Surrogate data for hypothesis testing of physical systems. *Phys. Rep.* 748: 1-60.

Lee R and Miller T 2002 An approach to forecasting health expenditures, with application to the U.S. medicare system. *Health Serv. Res.* 37(5): 1365-1386.

Meyers G 2005 On Predictive modeling for claim severity. In: *Casualty Actuar. Soc. Forum* 215-253.

Mwangi M and Murigu J 2015 The determinants of financial performance in general insurance companies in Kenya. *Eur. Sci. J.* 11(1): 288–297.

Pflaumer P 1992 Forecasting US population totals with the Box-Jenkins approach. *Int. J. Forecast.* 8(3): 329-338.

Renshaw AE 1994 Modelling the claims process in the presence of covariates. *ASTIN Bull. J. IAA.* 24(2): 265-285.

Tiao GC 1985 Autoregressive moving average models, intervention problems and outlier detection in time series. *Handb. Stat.* 5: 85–118.

Zheng A, Fang Q, Zhu Y, Jiang C, Jin F and Wang X 2020 An application of ARIMA model for predicting total health expenditure in China from 1978-2022. *J. Glob. Health* 10(1): 1-8.