Research Journal of Mathematical and Statistical Sciences _____ISSN 2320-6047
Vol. **10(1),** 1-8, January **(2022)** *Res. J. Mathematical and Statistical Sci.*

Prediction model for registered motor vehicles based on box-Jenkins approach

Paul Boye[1 *], Frederick Eshun[2] and Agyarko Kofi[1]
[1]Faculty of Engineering, Department of Mathematical Sciences, University of Mines and Technology, Tarkwa, Ghana
[2]College of Distance Education, University of Cape Coast, Accra Regional Office, Papafio Hills, Ghana
pboye@umat.edu.gh

Available online at: www.iscamaths.com , www.isca.in , www.isca.me
Received 27th January 2021, revised 11th May 2021, accepted 22nd September 2021

Abstract

The number of motor vehicles to be registered in a country is an important guiding standard for sustained economic growth. However, there are numerous challenges customers face during the registration exercise which has been given no scholarly attention. Here, Seasonal Autoregressive Integrated Moving Average (SARIMA) is proposed to provide future prediction of annual number of motor vehicles to be registered in Ghana has been developed. This study uses vehicles of all categories monthly registered dataset over five-years which was obtained from Driver and Vehicle Licensing Authority (DVLA) in Accra the capital city of Ghana to develop a SARIMA model by using Box-Jenkins approach for future prediction of motor vehicles to be registered at DVLA. In the modeling, the seasonality component in the dataset was taken care of by the process of differencing. The developed model performance was assessed based on good statistical indicators such as Mean Absolute Percentage Error (MAPE), Normalized Root Mean Square Error (NRMSE) and Relative Percentage Error (RPE). Results confirmed that the SARIMA model can be used to predict the number of motor vehicles to be registered annually in a country. This study is useful and a major contribution for modeling the expected number of motor vehicles of all categories to be registered in a country within the year.

Keywords: DVLA, vehicle registration, SARIMA, stock prediction, policymakers.

Introduction

The exponential growth of motor vehicle ownership and people in up-and-coming markets is a global phenomenon with little attention to how such an increase would affect traffic and pedestrian flow within most of the metropolitan cities in the world[1]. This situation has serious implications for transportation and environmental policies[2]. Numerous studies in literature on modeling traditional data have been carried out in various experimental fields such as stock prediction[3], hydrology[4], meteorology[5] and other related fields. The main superiority of the traditional models are simple computational process, high short-term forecast accuracy, and diagnostic checking.

Time series models like Seasonal Autoregressive Integrated Moving Average (SARIMA) can adequately explain the non stationary behaviours both within and across seasons in datasets[6]. In addition, the SARIMA model has the capability to predict future events adequately. Due to the strengths exhibited by SARIMA, the method has become the preferred choice for several time series modelers[7]. That is, the SARIMA prediction capabilities have been well documented in literature. Notable applied areas include aviation[8], medicine[9], energy[10] etc. Although its application is replete in literature, the SARIMA model's capability in predicting the number of motor vehicles to be registered annually in a country has been given no scholarly attention. Therefore, on the basis of the SARIMA reported advantages, it would be prudent to further expand their application frontiers. Hence, the research presents SARIMA as reliable tool that can be used to determine the number of motor vehicles to be registered. The propose SARIMA approach was tested on dataset acquired from Driver and Vehicle Licensing Authority (DVLA) of Ghana.

The DVLA is a government agency responsible for the licensing and evaluation of drivers and motor vehicles. It is not unusual that in every year new motor vehicles are imported into the country and registered at the various DVLA branches in the country. This means that, the importance of transportation to a nation's socio-economic growth cannot be underestimated.

Studies have shown that the Process time variability is a serious challenge customer face from the license vehicle number plate production plants in meeting up with daily demands and this causes long queues and waiting[11]. The number of registered motor vehicles is growing very rapidly in Ghana; whilst it is an indication of sustained economic growth, it presents a major challenge to the country's transport policymakers. The stupendous growth of vehicular population results in an instantaneous impact of individual income growth. This reveals a healthy relation between development in per capita income and it also implies welfare gain. But the existence of negative externalities like traffic congestion and air contamination shows wealth loss.

The registration policy is very important for a number of reasons, which include generation of revenue for economic development and checking of crime by using the motorised details[12].

According to Boah-Mensah[13] and Agunbiade and Peter[14] a solution to these predicaments would be achieved if service providers and policymakers will have comprehensive knowledge of the additional number of motor vehicles to be registered annually in a country. The present authors were therefore motivated to develop a prediction model that can provide in advance the number of vehicles expected to be registered by the DVLA. In that regard, time spent at DVLA registration centers could be reduced, because management could rely on the proposed model to make projections and that will help them to make well informed organisational decisions and proper arrangements to fast track the registration process. Consequently, management will achieve an acceptable customer service level. In addition, good infrastructure would also be put in place to mitigate the challenges pedestrians suffer.

Related works: For organizations to achieve credible operative performance, it is imperative to estimate in advance client's requests. According to Abu and Ismail[15], vehicle request estimation was in short supply due to lack of information. The authors fitted an appropriate vehicle estimation model using Box-Jenkins method based on nine years Malaysia dataset. Reliable quality model diagnostic checks were performed and Malaysian private vehicle demand was estimated. Using Minitab software, revealed results showed that SARIMA (2, 1, 0) (2, 0, 0) [12] was the most satisfactory model for estimating individual vehicle demand.

According to Anvari et al[16] a well-organized and properly running of government transportation systems is of great importance to the fast-growing civilizing world. However, estimating transportation request is vital for policymakers planning and supervision. In a case like this, the authors used traditional model based on Box-Jenkins technique for such estimation. Using MATLAB, the authors applied reliable statistical tests for a good model selection and it was tested on Istanbul Metro dataset. Experimental results revealed that the suggested model was good.

Data generation

During the model development, vehicles of all categories monthly registration dataset was obtained from Driver and Vehicle Licensing Authority (DVLA) Accra the capital city of Ghana for a period of five-years. Table-1 presents descriptive records of the Total Registered Motor Vehicles (TRMV).

Table-1: TRMV statistical records.

Parameter	Data Size	Max. Value	Min. Value	Average	Standard Deviation
TRMV	60	27 101	2 020	11 870	5 150.428

Overview of methods applied: Figure-1 shows the general methodology for the methods employed.

Development of non-seasonal ARIMA model: In literature, Kirchgässner et al[17] and Walter and Pascalau[18] have indicated that the Box-Jenkins Autoregressive Moving Average (ARMA) model is a combination of the Autoregressive and Moving Average models (Equation-1) respectively.

$$X_t = a_0 + \sum_{p=1}^{P} a_p x_{t-p} + \sum_{q=1}^{Q} b_q \varepsilon_{t-q} \quad (1)$$

where a_0, a_p and b_q are constant and model coefficient parameters respectively. *P* and *Q* are the number of observations and error terms respectively. x_{t-p} and ε_{t-q} are lagged observation and error terms at periods $t-p$ and $t-q$ respectively. The hypothesis governing the Box-Jenkins method is given as: i. H_0: The series is nonstationary. ii. H_1: The series is stationary.

Seasonal ARIMA model: According to Hyndman and Athanasopoulos[19], SARIMA models (Equation-2) are obtained from Equation-1 when non-seasonal terms are added to it.

$$\text{SARIMA}\ (p,\ d,\ q)\ (P,\ D,\ Q)\ [\text{m}] \quad (2)$$

here m = observations per year.

It can be formulated as follows: Let B = Backshift operator.

$$Bx_t \equiv x_{t-1}$$
$$B^2 x_t = BBx_t = Bx_{t-1} = x_{t-2}$$
$$\vdots$$
$$B^d x_t = x_{t-d}$$

Let ∇ be non-seasonal differencing operator.

Define $\nabla \equiv (1-B)$

$$\nabla x_t = (1-B)\, x_t = x_t - x_{t-1}$$
$$\nabla^2 x_t = (1-B)^2\, x_t = (1-2B+B^2)\, x_t = y_t - 2x_{t-1} + x_{t-2}$$
$$\vdots$$
$$\nabla^d x_t = (1-B)^d\, x_t = \left(\sum_{i=0}^{d} \binom{d}{i} (-B)^{d-i}\right) x_t \text{; from the Binomial Theorem.}$$

Let ∇_s be seasonal differencing operator (s is season period).

Define $\nabla_s \equiv (1-B^s)$

$$\nabla_s x_t = (1-B^s)\, x_t = x_t - x_{t-s}$$

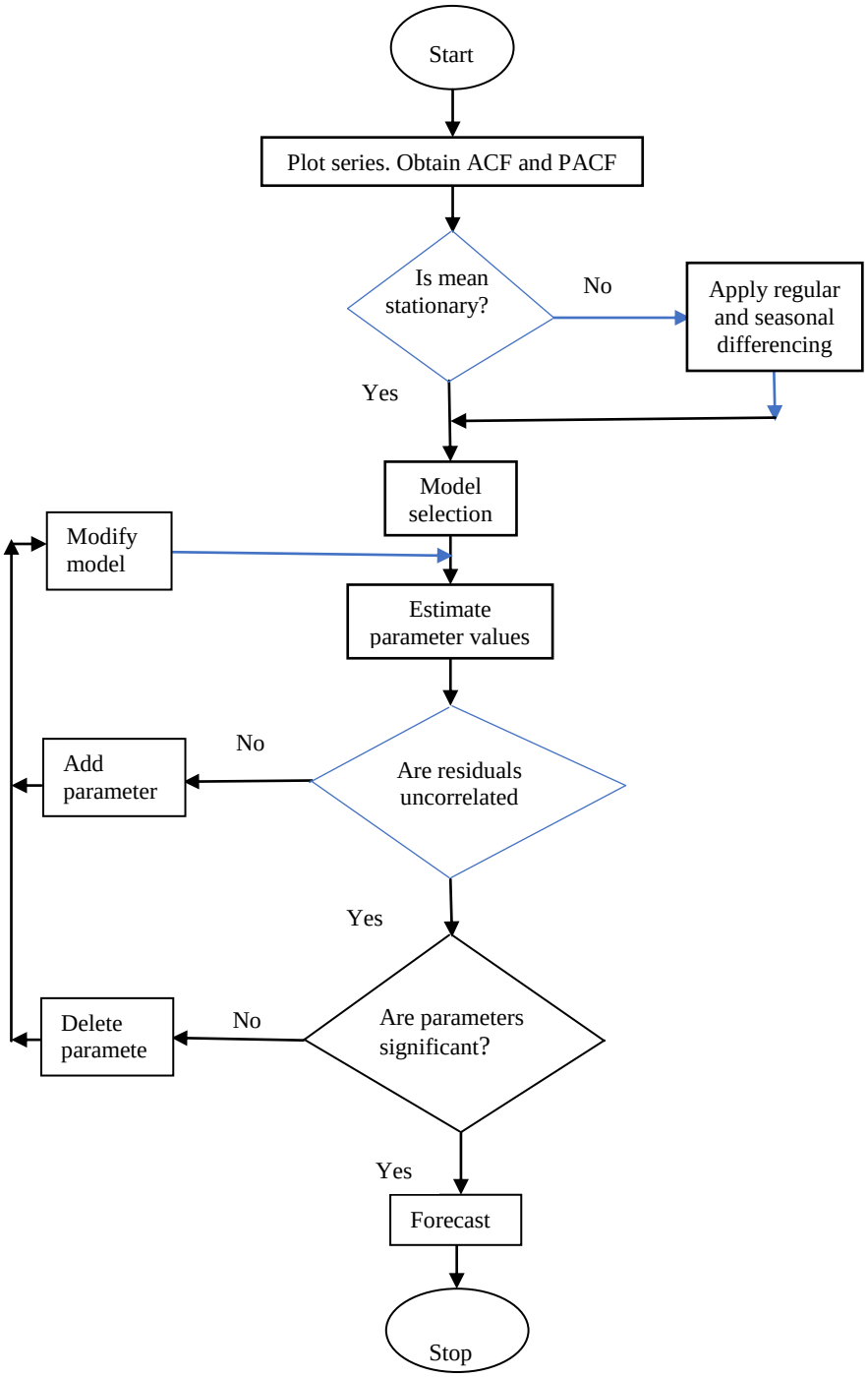

Figure-1: General Methodology.

$$\nabla_s{}^2 x_t = \left(1-B^s\right)^2 x_t = \left(1-2B^s+B^{2s}\right)x_t = x_t - 2x_{t-s} + x_{t-2s}$$

$$\vdots$$

$$\nabla_s{}^D x_t = \left(1-B^s\right)^D x_t = \left(\sum_{i=0}^{D}\binom{D}{i}\left(-B^s\right)^{D-i}\right)x_t$$

Non-seasonal Autoregressive Model, AR (p)

$$x_t = a_0 + \sum_{i=1}^{p} a_i x_{t-i} + \varepsilon_t$$

$$\left(1-a_1B-a_2B^2-\ldots-a_pB^p\right)x_t = a_0+\varepsilon_t$$

$$\phi_p\left(B\right)x_t = a_0+\varepsilon_t$$

where $\phi_p(B) = \left(1 - a_1 B - a_2 B^2 - ... - a_p B^p\right)$

Seasonal Autoregressive Model, SAR (p)

$$x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + \varepsilon_t$$

$$\left(1 - A_1 B^s - A_2 B^{2s} - ... - A_p B^{ps}\right) x_t = A_0 + \varepsilon_t$$

$$\Theta_p\left(B^s\right) x_t = A_0 + \varepsilon_t$$

where $\Theta_p\left(B^s\right) = \left(1 - A_1 B^s - A_2 B^{2s} - ... - A_p B^p\right)$

Non-seasonal Moving Average Model, MA (q)

$$x_t = b_0 + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t$$

$$= b_0 + \left(1 - b_1 B - b_2 B^2 - ... - b_q B^q\right) \varepsilon_t$$

$$= b_0 + b_q\left(B\right) \varepsilon_t$$

where Seasonal Moving Average Model, SMA (Q)

$$x_t = B^s{}_0 + \sum_{j=1}^{Q} B^s{}_j \varepsilon_{t-j} + \varepsilon_t$$

$$= B^s{}_0 + \left(1 - B_1 B^s - B_2 B^{2s} - ... - B_Q B^{Qs}\right) \varepsilon_t$$

$$= B_0 + B_Q\left(B^s\right) \varepsilon_t$$

where $\Phi_Q\left(B^s\right) = \left(1 - B_1 B^s - B_2 B^{2s} - ... - B_Q B^{Qs}\right)$

Non-seasonal Differencing Model, I(d)

$$\nabla^d x_t = a_0 + \varepsilon_t$$

$$\left(1 - B\right)^d x_t = a_0 + \varepsilon_t$$

Seasonal Differencing Model, SI(D)

$$\nabla_s{}^D x_t = b_0 + \varepsilon_t$$

$$\left(1 - B^s\right)^D x_t = b_0 + \varepsilon_t$$

Therefore $\text{ARIMA}\left(p,\, d,\, q\right)\left(P, D, Q\right)$ model is given by

$$\phi_p\left(B\right) \quad \Theta_p\left(B^s\right) \quad \nabla^d \quad \nabla_s{}^D x_t \;=\; b_0 + \quad \varphi_q\left(B\right) \Phi_Q\left(B^s\right) \varepsilon_t$$

Model Assessment: The developed model performance was accessed by using Mean Absolute Percentage Error (MAPE) (Equation-3), Normalized Root Mean Square Error (NRMSE) (Equation-4) and Relative Percentage Error (RPE) (Equation-5)[20-23].

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\varepsilon_i}{x_i} \right| \times 100\% \tag{3}$$

$$NRMSE = \frac{RMSE}{x_{\max} - x_{\min}} \tag{4}$$

$$RPE = \sum_{i=1}^{N} \frac{\left|\varepsilon_i\right|}{x_i} \times 100\% \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \varepsilon_i{}^2} \,. \tag{6}$$

where ε_i is the i^{th} error, x_i is the i^{th} data, N is data size, TRMV = total registered motor vehicles.

Results and discussion

Data pretreatment: Table-1 indicates the initial analysis of the TRMV dataset using r statistical software version 3.6.1. The standard deviation of 5150.428 (Equation-7) reveals that the registered motor vehicles dataset is widely spread out from the mean value of 11 870 (Equation-8).

$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2} \tag{7}$$

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N} \tag{8}$$

The dataset was tested for stationary by plotting time series, ACF and PACF graphs. From the graphical results, the dataset was found to be nonstationary. Consequently, it was partitioned into 55 (approximately 91.6666%) training set and 5 (approximately 8.3333%) testing set to examine the strength and accuracy of the trained model. The training set which was used for the SARIMA modeling was subjected for stationary test by plotting time series graph and ACF and PACF graphs. From the results of the graphs and a formal Augmented Dickey-Fuller test revealed that the trained set was not stationary. As a result, it was differenced once and it was found to be stationary. Hence, the SARIMA model was developed based on the stationary trained set.

The SARIMA model: The TRMV series (Figure-2) wanders up and down and shows some seasonal components. The ACF and PACF plot (Figure-3) decayed slowly. Clearly, the series can be said to be nonstationary.

Similarly, the trained set (Figure-4) which was used to develop the ARIMA model wanders up and down and shows some seasonal components which made it to be nonstationary. The ACF and PACF (Figure-5) decayed slowly respectively. A formal stationary test, Augmented Dickey-Fuller test was also performed (Table-2) and it has a value of -3.45 at lag 3. Clearly, from the p-value = 0.05723 > 0.05 is an indication of accepting the null hypothesis that the training set is not stationary. Consequently, it was differenced to take care of the seasonality components from the dataset.

Table-2: Formal stationary test.

Augmented Dickey-Fuller Test	Lag order	P-value
-3.45	3	0.05723

The trained series time plot (Figure-6) clearly wanders up and down around the zero-mean line. This is an indication that the trained series seemed to be stationary after the first difference. The series first difference ACF plot which is used to obtain the order q value for the moving average component shows significant spikes at lags 0, 0.1, 0.9, 1 and 1.1 respectively (Figure-7). The corresponding PACF plot which is used to obtain the order *p* value for the autoregressive component decays slowly. A formal test, Augmented Dickey-Fuller test (Table-3) was also performed; it had a value of -4.883 at lag 3 which is more negative compared to the former. Thus, more negative the test value the better the stationary condition. A p-value of 0.01 which is less than 0.05 is a clear indication that the train series is stationary.

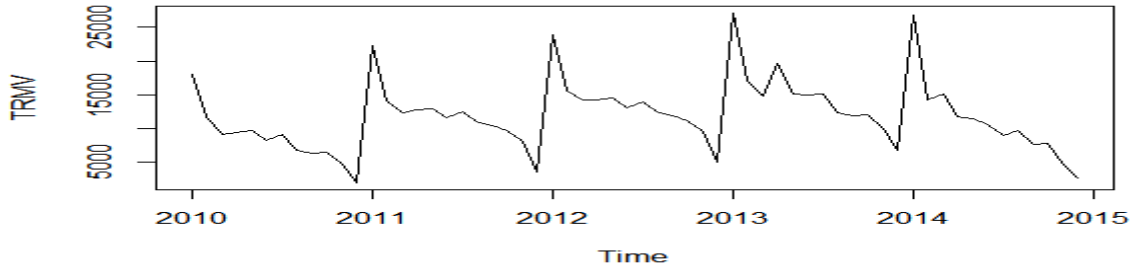

Figure-2: Time Series Plot of Total Registered Motor Vehicles.

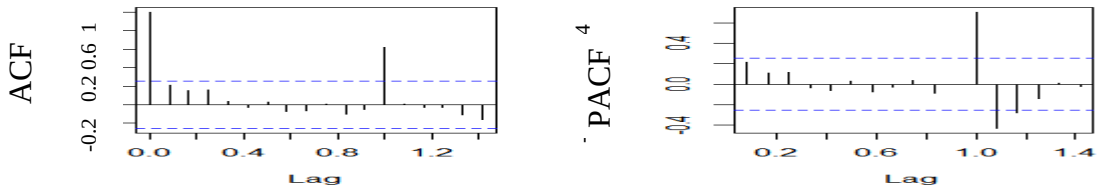

Figure-3: ACF and PACF of Time Series of Total Registered Motor Vehicles.

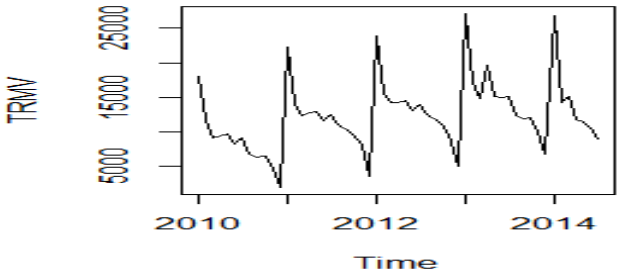

Figure-4: Trained Series of Registered Motor Vehicles.

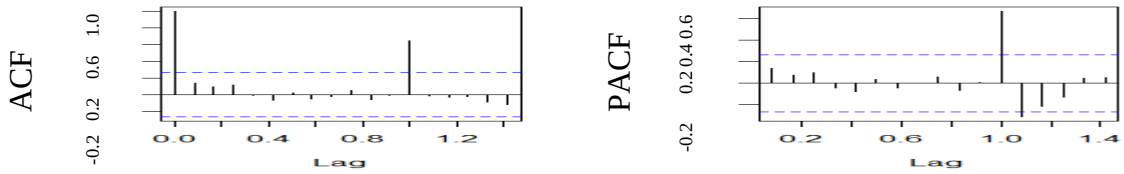

Figure-5: Trained Series ACF and PACF.

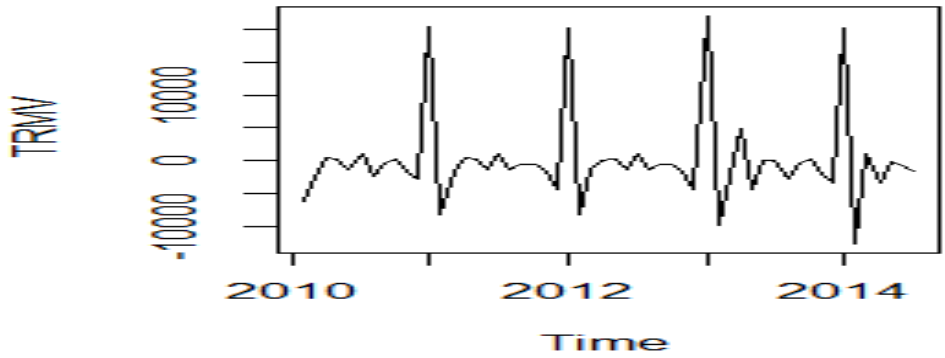

Figure-6: Trained Series First Difference.

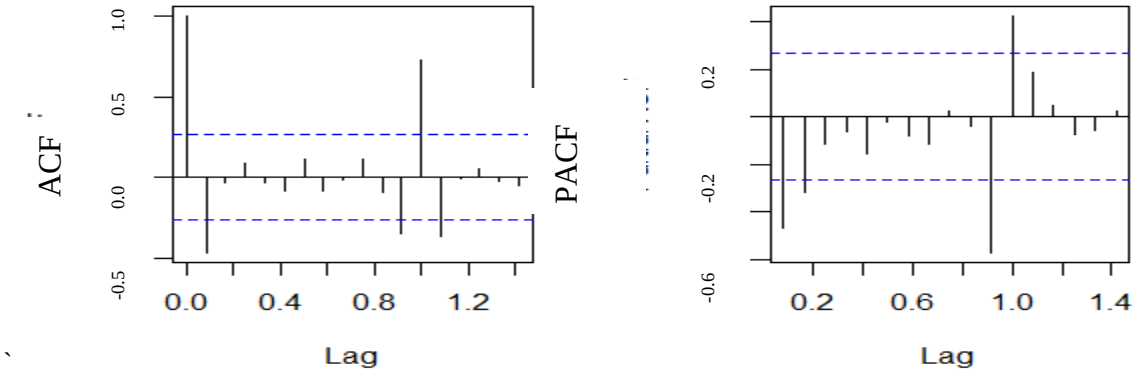

Figure-7: First Difference ACF and PACF.

Table-3: Formal test of first difference.

Augmented Dickey-Fuller Test	Lag order	P-value
-4.883	3	0.01

Table-4 shows the possible models which were obtained when the trained series was analysed and run by R software version 3.6.1. Consequently, the best SARIMA model was obtained as SARIMA (0, 1, 2) (0, 1, 1) ^{12} based on the minimum value of AIC as 733.6383. The seasonal component of (0, 1, 1) was also obtained from the results of the R-software.

Table-4: Model selection criteria.

Candidate Model	AIC
SARIMA (2, 1, 2) (1,1,1) ^{12}	737.9769
SARIMA (1, 1, 0) (1,1,0) ^{12}	739.7066
SARIMA (0, 1, 2) (0,1,1) ^{12}	733.6383
SARIMA (1, 1, 2) (0,1,1) ^{12}	735.1898
SARIMA (2, 1, 2) (1,1,0) ^{12}	735.4095

Developed model efficiency test: The statistical indicators (MAPE, NRMSE, RPE) in Table-5 for the testing set were invoked to examine the strength and accuracy of the trained SARIMA (0, 1, 2) (0, 1, 1) ^{12} model. The MAPE value of 10.1375 shows that the developed model was able to explain 89.8625% variability in the test data. For RPE value of 50.6873%, it can be said that the developed model revealed 49.3127% prediction precision. But on the other hand, NRMSE value of 0.08818 portrays the amount of error the developed model could not explain. Consequently, all these outcomes show that the developed model is good.

Table-5: Results of statistical performance indicators.

Model	MAPE	Testing Set	
		NRMSE	RPE
SARIMA (0, 1, 2) (0,1,1) ^{12}	10.1375	0.08816	50.6873

Conclusion

In this study, SARIMA model has been developed from vehicles of all categories monthly registered dataset over five-years which was obtained from DVLA in Accra the capital city of Ghana, to predict the number of vehicles to be registered in the country annually. In developing the model, dataset nonstationary which could have resulted in erroneous results was resolved by train data differencing.

The resultant SARIMA model which explained 89.8625% variability in the testing set is SARIMA (0, 1, 2) (0, 1, 1) ^{12}. The RPE value of 50.6873% shows that the developed model measured 49.3127% estimation precision. For NRMSE value of 0.08818, is an indication of error margin the developed model could not explain, meaning that the developed model is good.

The approach presented in this study for deriving the SARIMA model to estimate the TRMV is a valuable contribution to the body of knowledge in modeling which should be adopted by the DVLA of Ghana and countries with similar registration problems to estimate the expected number of motor vehicles to be registered annually.

Acknowledgement

The authors gratefully acknowledge DVLA officials of Ghana for providing the data to make this study a successful one.

References

1. Kumar, S.V. & Vanajakshi, L. (2015). Short-Term Flow Prediction Using Seasonal ARIMA Model with Limited Input Data. *European Transport Research Review*, 7, 1– 9.

2. Dargay, J., Gately, D. & Sommer, M. (2007). Vehicle Ownership and Income Growth. *Worldwide*: 1960-2030, 1-30.

3. Wahyudi, S.T. (2017). The ARIMA Model for Indonesia Stock Price. *International Journal of Economics and Management*, 11(1), 223-236.

4. Katimon, A., Shahid, S., & Mohsenipour, M. (2018). Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia. Sustainable Water Resources Management, 4(4), 991-998. https://doi.org/ 10.1007/s40899-017-0202-8.

5. Murat, M., Malinowska, I., Gos, M. & Krzyszczak, J. (2018). Forecasting Daily Meteorological Time Series Using ARIMA and Regression Models. *Journal of International Agrophysics*, 32, 253-264. https://doi.org/10.1515/intag-2017-0007.

6. Wang, S., Feng, J. & Liu, G. (2013). Application of Seasonal Time Series Model in the Precipitation Forecast. *Mathematical and Computer Modelling*, 58, 677-683.

7. Afrifa-Yamoah, E. & Saeed, B.I.I. (2016). Sarima Modeling and Forecasting of Monthly Rainfall in the Brong Ahafo Region of Ghana. *World Environment*, 6(1), 1-9. https://doi.org/10.5923/j.env.20160601.01.

8. Ruiz-Aguilar, J., Turias, J. & Jiménez-Come, M.J. (2014). Hybrid Approaches Based on SARIMA and Artificial Neural Networks for Inspection Time Series Forecasting. *Transportation Research Part E*, 67, 1-13. http://dx.doi.org/10.1016/j.tre.2014.03.009.

9. Qi, C., Zhang, D., Zhu, Y., Liu, L., Li, C., Wang, Z. & Li, X. (2020). SARFIMA Model Prediction for Infectious Diseases: Application to Hemorrhagic Fever with Renal Syndrome and Comparing with SARIMA. *BMC Medical Research Methodology*, 20(243), 1-7. https://doi.org/10.1186/s12874-020-01130-8.

10. Suksawang, P., Suphachan, S. & Kaewnuch, K. (2018). Electricity Consumption Forecasting in Thailand Using Hybrid Model SARIMA and Gaussian Process with Combine Kernel Function Technique. *International Journal of Energy Economics and Policy*, 8(4), 98-109.

11. Phamotse, I.M. & Lues, L. (2010). Assessing Practices During the Vehicle Registration Process in Lesotho. *Journal for New Generation Sciences*, 8(1), 190-203.

12. Le Vine, S., Wu, C., & Polak, J. (2018). A nationwide study of factors associated with household car ownership in China. IATSS research, 42(3), 128-137.

13. Boah-Mensah, E. (2013). The Number of Cars in Ghana Increase by 23%. *www.vehicle population in Ghana*. (Accessed March 19, 2015).

14. Agunbiade, D.A. & Peter, E.N. (2013). Modeling and Forecasting Vehicle Registration System: An ARMA Approach. *Society for Mathematical Services and Standards*, 2(1), 1-13.

15. Abu, N. & Ismail, Z. (2019). A Study on Private Vehicle Demand Forecasting Based on Box-Jenkins method. *AIP Conference Proceedings 2059*, 1-9. AIP Publishing LLC. https://doi.org/10.1063/1.5085948.

16. Anvari, S., Tuna, S., Canci, M. & Turkay, M. (2016). Automated Box-Jenkins Forecasting tool with an Application for Passenger Demand in Urban Rail Systems. *Journal of Advanced Transportation*, 50, 25-49. https://doi.org/10.1002/atr.1332 .

17. Kirchgässner, G., Wolters, J. & Hassler, U. (2013). Univariate Stationary Processes. Introduction to Modern Time Series Analysis, Springer, 27-93.

18. Walter, E. & Pascalau, R. (2015). Pretesting for multi-step-ahead exchange rate forecasts with STAR models. *International Journal of Forecasting*, 31(2), 473-487.

19. Hyndman, J.R. & Athanasopoulos, G. (2018). Forecasting: Principles and Practice.

20. Aman, S., Simmhan, Y., & Prasanna, V. K. (2014). Holistic measures for evaluating prediction models in smart grids. IEEE Transactions on Knowledge and Data Engineering, 27(2), 475-488.

21. Wang, C. N., & Phan, V. T. (2014). An improvement the accuracy of grey forecasting model for cargo throughput in international commercial ports of Kaohsiung. International Journal of Business and Economics Research, 3(1), 1-5.

22. Chen, C. & Hsin, P. (2016). Forecasting of Taiwan's Gross Domestic Product Using the Novel Nonlinear Grey Bernoulli Model with ANN Error Correction. *Journal of Global Economics*, 4, 1-4.

23. Moonchai, S. & Rakpuang, W. (2015). A New Approach to Improve Accuracy of Grey Model GMC (1, n) in Time Series P. *Modelling and Simulation in Engineering*, 1-10. http://dx.doi.org/10.1155/2015/126738