Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.10, 2015

ANALYZING THE IMPACT OF HISTORICAL DATA LENGTH IN NON SEASONAL ARIMA MODELS FORECASTING

Amon Mwenda, Dmitry Kuznetsov, Silas Mirau
School of Computational and Communication Science and Engineering
Nelson Mandela Institution of Science and Technology (NM-AIST), P.O. box 447, Arusha, Tanzania

Abstract

Different values of minimum data requirement for ARIMA models have been proposed. It also proposed to use as much data as they are available in formulating ARIMA models. This paper studied the impact of the size of the historical data on ARIMA models in forecasting accuracy. The study used 286 weekly records of amount of solid waste generated in Arusha City to formulate four ARIMA models using different data lengths or size. The first model, M1 used 30 observations, the second model, M2 used 60 observations, the third model M3 used 120 observations and the fourth model, M4 used 260 observations all of which are the most recent. A total of 26 observations were held out for validation. The precision in forecasting was tested using MAPE, RMSE and MAD. The results indicated variation in precision. M3 performed best in one-week ahead and 9 – 12 weeks ahead while M4 did best in 2 – 8 weeks and also for 13 weeks and above. M1 was the worst model in forecasting.

Keywords: ARIMA models, MAPE, RMSE, MAD, Forecasting

1. Introduction

Time series ARIMA models are popular in forecasting univariate historical data. Despite the popularity, there are still some areas that have been given little attention in the literature. One area is the length of historical observations required for an ARIMA model to produce forecast with high precision. According to literature, Box and Jenkins (1976) who are pioneers of ARIMA Box-Jenkins modeling approach recommended a minimum of 40 or 50 past observations to meet the objective of obtaining a sufficiently good forecasting with a narrow prediction interval. According to Hyndman and Kostenko (2007) and Chatfield (1996), the number of observations in any statistical model rely on the number of parameters and the random variation in the data. They argued that number of observations should be greater than the number of parameters to be estimated and they should increase accordingly for data with a lot of random variations. They recommended that for a seasonal ARIMA model with 15 parameters, number of observations should be at least 16. Hanke (1998) proposed a table with minimum number of observations for various models, however, it was criticized for overlooking random variation in data. There is no maximum limit set but to use as much data as they are available. Does accuracy in forecasting increases with increased data? In the literature, different lengths of data have been used to formulate models. Some forecasters have used as less as 20 while others have used quite big number of past data. None of the authors have stated reasons for the number of past data they have chosen to use. Sarpong (2013) conducted a study on maternal mortality forecasting in Ghana with quarterly 50 observations and concluded that $ARIMA(1,0,2)$ adequately fitted the data than others. The model was used to forecast 20 quarterly periods ahead. Biswas and Bhattacharyya (2013) used 57 observations to build ARIMA models and only 3 observations for model testing in a study to forecast area and production of rice in West Bengal and the models exhibited good accuracy in forecasting. Alba and Mendoza (2007) in their study of forecasting methods for short time series, used 24 monthly observation to formulate ARIMA model although they concluded that it was not possible to identify some of the time series components such as seasonality with such small observations even when the series is known to be seasonal. Paul, Hoque & Rahman (2013) examined empirically ARIMA model to forecast average daily share price index of pharmaceutical companies in Bangladesh used a total of 236 observations from Dhaka stock exchange (DSE). In this study, $ARIMA(2,1,2)$ was found to be the best model. Another study to forecast road accident injuries in Ghana used only 21 annual injuries records from 1991 to 2010 to formulate the best fit $ARIMA(1,1,1)$ that was used to forecast one year period ahead producing 13337 accidents against observed 13272 injuries for the year 2011 (Ofori, Ackah and Ephraim, 2013). A study to assess and forecast the contribution of industrial sector to GDP of Bangladesh used only 33 observations of yearly industrial contribution records from 1979/1980 to 2010/2011 in which 28 observations were used to develop

ARIMA model and 5 observations were used as out of sample test (Khan and Rahman, 2013). A study forecasting inflow of Dokan reservoir used over 600 records of monthly water inflow from 1953/1954 to 2004/2005 to formulate ARIMA model. The formulated seasonal model was used to forecast inflow from 2005 to 2007 (Al-Masudi, 2011). Another study on analysis of gas prices for Turkey from 2003 to 2011 used 36 monthly price observations to develop ARIMA and Exponential smoothing models in which $ARIMA(1,1,0)$ was found to be the most efficient model and was used to forecast gas price for Turkey for the next 12 periods ahead (Wilberforce, 2013). A study of forecasting municipal solid waste generation in Kumasi Metropolitan Area used 72 observations of monthly solid waste generation from January 2005 to December 2010. The most efficient model was $ARIMA(1,1,0)$ which was used to forecast solid waste generation for 36 monthly periods ahead up to December 2013 (Ebenezer O., Emmanuel & Ebenezer B., 2013). A study to forecast paddy cultivated area and paddy production in Bastar division of Chhattisgarh used 36 annual observations. The results showed that $ARIMA(2,1,2)$ and $ARIMA(2,1,0)$ fitted best for forecasting cultivated area and production level respectively (Singh, Kumar & Prabakaran 2013). From the surveyed literature, ARIMA models were formulated using between 24 and 600 actual observations.

This paper is analyzing the impact of the length of the observed data in formulating ARIMA models and selection of best model based on the accuracy of prediction from the models. Section two explains the methodology of ARIMA models, section 3 is about the four formulated models using most recent data of different lengths, section 4 is results and discussion and section 5 is conclusion and recommendations.

2. ARIMA Models and analysis procedure

Four ARIMA models with varied number of observations are specified. The 287 observations of weekly solid waste generation from July 2008 to December 2013 are used. 27 data values are hold for out-of-sample testing purposes. The remaining data values are apportioned into four subgroups each of which is used to formulate an ARIMA model. The first model uses 25 observations, the second model uses 30 observations, the third model uses 60 observations and the final model uses 180 observations. The models are formulated by the Box-Jenkins methodology which has basic four stages; Stationarity checking, model identification, parameter estimation and diagnostic checking.

Stationarity checking

Stationarity of the observed data are checked by means of time series plots observations and unit root tests. If the data exhibits non stationarity properties, then it is transformed by first differencing to obtain a stationary series and needed second differencing is taken.

Identification

Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) plots of stationary series are examined to identify the orders of the autoregressive and moving average parameters of the ARIMA model to be formulated. The order of autoregressive part is given by the lag at which PACF cuts off to zero and the order of the moving average is given by the lag at which ACF cuts off to zero.

Estimation

Different parameters estimation methods exist in the literature. The method of exact maximum likelihood estimation is used to estimate parameters by means of gretl statistical software. Since the objective of this paper is to analyze the impact of number of observations used in developing the ARIMA model, parameter standard errors for each model is computed.

Diagnostic checking

Correlogram of the residuals of the formulated models are checked for significance. If they are not significantly different from zero, the models are assumed to adequately fit the data. Additionally, Ljung-Box test statistic or Q - statistic is computed.

Analysis procedure

To analyze the impact of observations size used in formulating ARIMA models, four models using 30, 60, 120 and 260 weekly observations are formulated using Box-Jenkins approach. Performance in terms of MAPE, RMSE and MAD is computed and compared.

3. Data and models formulation

Data used are weekly observation of solid waste generation for a period of five years. Four ARIMA models, M1, M2, M3, and M4 are considered.

M1 model: The first model M1 has 30 most recent weekly observations. The time series plot for M1 does not

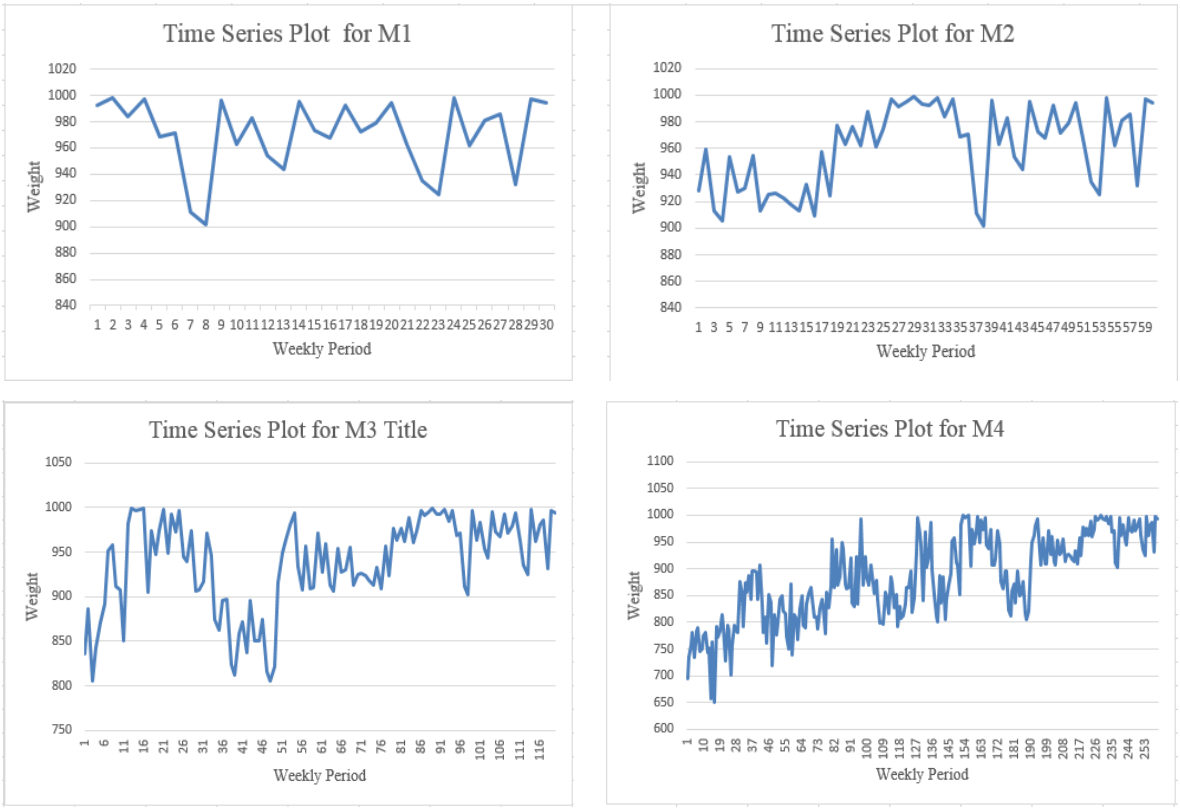
exhibit any trend. The ADF test has p-value less than the 0.05 significant level indicating that the series is stationary. The ACF and PACF plots do not contain any significant lags for determining the order of the model and do not give indicator of the order of ARMA model therefore several ARMA model are formulated and compared based on minimum Akaike information criterion (AIC).

Table 1. AIC for the Formulated ARMA Models

Order (p/q)	0	1	2
0		287.43	285.55
1	287.52	285.09	286.89
2	289.19	287.17	285.52

So $ARMA(1,1)$ is found to have minimum AIC among the tested models as shown in table 1. Diagnostic checking show that residuals are random, so $ARMA(1,1)$ is the best fit for 30 observations.

Figure 1. Time Series Plots for the Four Models

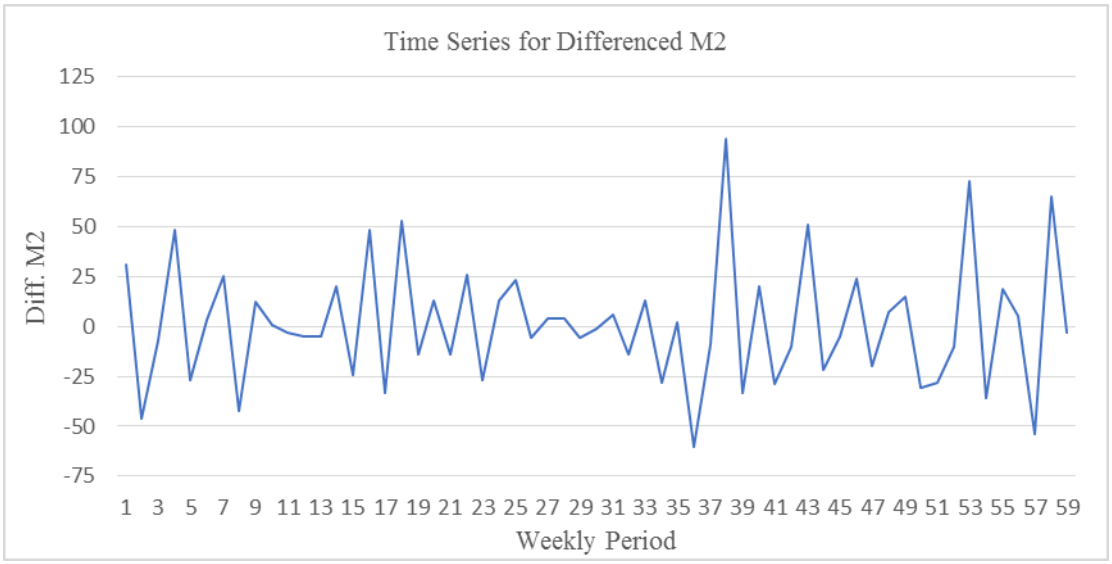


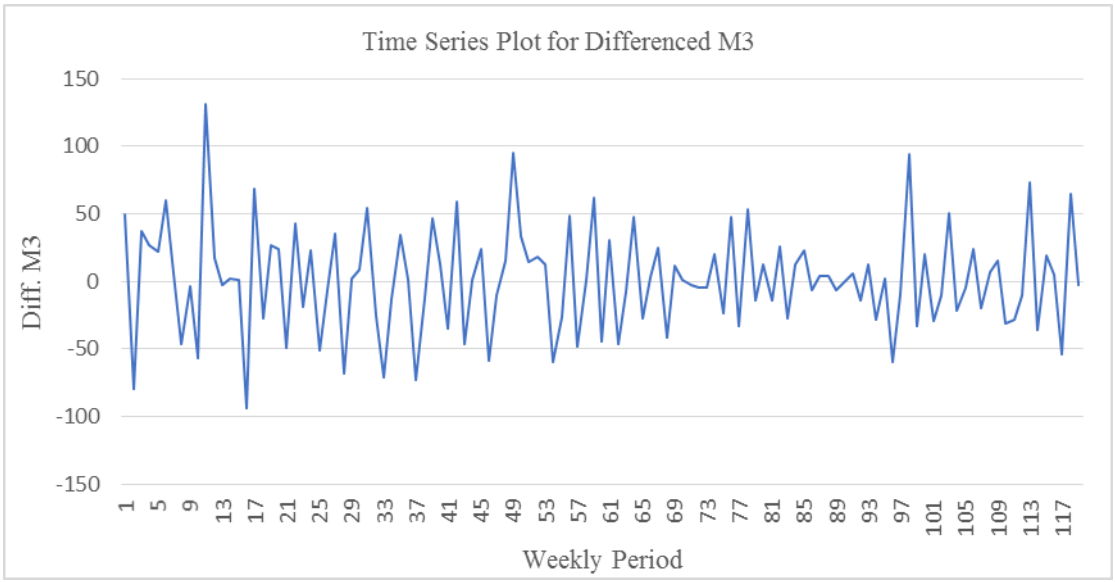
M2 model: The second model M2 has 60 most recent weekly observations. The time series plot for M2 exhibits a weak trend with fluctuations suggesting that the series is not stationary in both mean and variance. The series is first differenced and the plot of differenced series looks stationary about zero mean in figure 2. The first differenced series looks stationary about the zero mean. This is confirmed by ADF test which yield a p-value of 0.0000000154 which is much less than the 0.05 critical value. Hence the series is stationary. Both ACF cut off at lag 1 but lag 4 is also significant and PACF cuts off at lag 2 but lag 4 is also significant. The proposed models are $ARIMA(2,1,0)$; $ARIMA(0,1,1)$, $ARIMA(4,1,0)$ and $ARIMA(0,1,4)$. For diagnostic checking, Box Ljung test statistics for all models are less than their respective critical values hence the residuals are random. Based on AIC, $ARIMA(0,1,1)$is the best fit for the 60 observations.

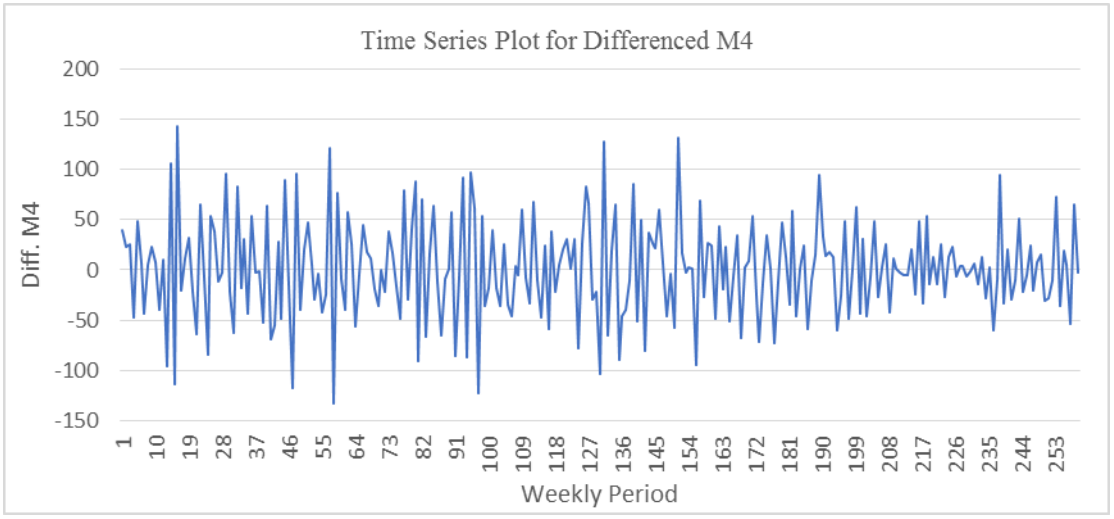
M3 Model: The third model M3 has 120 most recent weekly observations. The time series plot for M3 indicate a visible upward and downwards slight trends. The computed ADF p-value of 0.2202 is greater than the critical

value of 0.05 hence we cannot reject the null hypothesis that there is unit root and the series is non-stationary. The first difference is taken and the time series plot for the differenced series in figure 2 looks stationary about the zero mean. The ACF cuts off at lag 1 and lag 4 is also significant and PACF of the cuts off at lag 2 and lag 4 is also significant. Hence proposed models are $ARIMA(2,1,0)$, $ARIMA(0,1,1)$, $ARIMA(4,1,0$ and $ARIMA(0,1,4)$. The Box-Ljung statistics for all models show that the residuals are random. Based on AIC, $ARIMA(0,1,1)$ is selected as the best fit for the data.

Figure 2. Time Series Plots for Differenced M2, M3 and M4







M4 Model: The fourth model M4 has 260 weekly observations. The time series plot for M4 exhibits an obvious trend confirming non stationarity property of the series. No seasonal patterns observed. The series is first differenced and the differenced time series plot in figure 2 looks stationary around zero mean. The computed ADF p-value is much less than the critical value of 0.05 hence the series is stationary. The ACF plot cut off at lag 1 although lag 4 is significant and the PACF cuts off at lag 2 and lag 5 is also significant. Hence the proposed models are $ARIMA(2,1,0)$, $ARIMA(0,1,1)$, $ARIMA(5,1,1)$ and $ARIMA(0,1,4)$. Based on AIC, $ARIMA(0,1,4)$ is the best fit for the 260 weekly observations.

4. Results and discussion

The four models considered are M1 – $ARMA(1,1)$ for 30 weekly observations; M2 – $ARIMA(0,1,1)$ for 60 weekly observations; M3 – $ARIMA(0,1,1)$ for 120 weekly observations and M4 – $ARIMA(0,1,4)$ for 260 weekly observations. The models validation is done by residual correlogram analysis and computed Box-Ljung Q-statistic test and are all found to adequately fit the data. SPSS was used to compute these values as shown in table 1.

Table 2. Ljung Box Q – Statistic for the Models

Model	Number of past observations used	Ljung-Box Q-Statistic	DF	Chi-Square Critical Values at 95%	p-value
M1	30	17.7	16	27.587	0.343
M2	60	15.8	17	27.587	0.537
M3	120	10.9	17	27.587	0.861
M4	260	22.3	14	23.685	0.072

The computed Q-statistics are greater are less than the critical values of the Chi – Square distribution for the given degrees of freedom, the p-values are all greater than the critical value of 0.05, hence there are no evidence to reject the null hypothesis that the residuals do not exhibit lack of fit hence they are white noise series.

Parameters for each model have been estimated by using exact maximum likelihood method using GRETL software and the equations of the models are adapted from the general $ARIMA(p,d,q)$ equation:

$$(1-\phi_1 B-\ldots-\phi_p B^p)(1-B)^d X_t = (1-\theta_1 B-\ldots-\theta_q B^q)a_t$$

Where p and q are the orders of autoregressive and moving average components respectively, B is the backshift operator, X_t is the quantity predicted at time t, ϕ_p and θ_q are parameters. Since all the models are first differenced and with only moving average component, the above equation is reduced to

$$X_t = X_{t-1} + \phi_1 X_{t-1} - \phi_1 X_{t-2} + a_t - \theta_1 a_{t-1} + c$$

Computed parameters for M1 are $\phi_1 = -0.616399, \theta_1 = 1.0000 \; and \; c = 970.32$

$$X_t = 970.32 - 0.6164X_{t-1} + a_t - a_{t-1} \quad (1)$$

Computed parameters for M2 are $\phi_1 = 0, \theta_1 = -0.683129 \; and \; c = -0.860120$

$$X_t = X_{t-1} + a_t + 0.6831a_{t-1} - 0.8601 \quad (2)$$

Computed parameters for M3 are $\phi_1 = 0, \theta_1 = -0.473563 \; and \; c = 1.17595$

$$X_t = X_{t-1} + a_t + 0.4736a_{t-1} + 1.1760 \quad (3)$$

Computed parameters for M4 are $\phi_1 = 0, \theta_1 = -0.558939, \theta_2 = -0.0136488, \theta_3 = -0.00142579, \theta_4 = -0.175941 \; and \; c = 0.950247$

$$X_t = X_{t-1} + a_t + 0.5589a_{t-1} + 0.0136a_{t-2} + 0.0014a_{t-3} + 0.1759a_{t-4} + 0.9502 \quad (4)$$

The models formulated are used to point forecast the next 26 weeks ahead and the results of each model is compared to the 26 actual observations held out during parameters estimation period. Table 2 indicate the observed and the predicted values by the four models. These values are used to compute MAPE, RMSE and MAD for the four models. The results for the first week ahead, twelfth week and twenty fourth week are examined. Table 3 gives the MAPE, RMSE and MAD for the four models over the intervals selected. From the table, M3 has minimum error in the first and twelfth weeks, it is therefore the best model to forecasting one – step and twelfth – step ahead. M4 has minimum error in the twenty fourth step ahead and therefore it is the best model to use in forecasting at that point.

To study the variation in details, mean absolute percentage error (MAPE) and root mean squared error (RMSE) for all the 26 ahead points are computed and plotted in figure 2. The figure shows that M3 is the best model in one step ahead forecast and M4 is the best model to forecast two to eight weeks ahead. To forecast 9 to 12 weeks ahead, M3 gives the better results. To forecast 13 weeks and above M4 produce better predictions. For the models formulated, there is no direct relationship between the length of past data and forecasting accuracy although models formulated with 60, 120 and 260 observations performed relatively better than 30 observations model.

Table 3. Observed and Forecasted Values by the Four Models

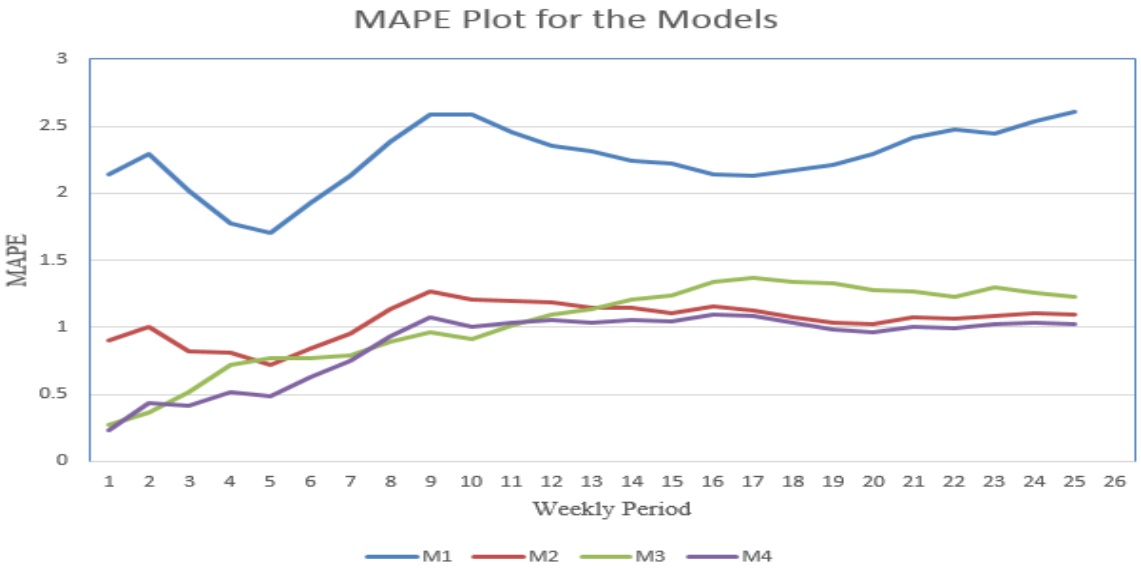
S/N	Observed Values	Predicted Values				S/N	Observed Values	Predicted Values			
		M1	M2	M3	M4			M1	M2	M3	M4
1	989	969.90	982.12	988.20	984.86	14	988	970.32	993.30	1003.48	995.29
2	994	970.57	982.98	989.37	993.43	15	983	970.31	994.16	1004.66	996.24
3	996	970.16	983.84	990.55	987.84	16	989	970.32	995.02	1005.84	997.19
4	982	970.41	984.70	991.72	985.79	17	978	970.31	995.88	1007.01	998.14
5	978	970.26	985.56	992.90	986.74	18	990	970.32	996.74	1008.19	999.10
6	984	970.35	986.42	994.08	987.69	19	1000	970.31	997.60	1009.36	1000.05
7	1003	970.29	987.28	995.25	988.64	20	1000	970.31	998.46	1010.54	1001.00
8	1006	970.33	988.14	996.43	989.59	21	1008	970.31	999.32	1011.72	1001.95
9	1015	970.31	989.00	997.60	990.54	22	1022	970.31	1000.18	1012.89	1002.90
10	1015	970.32	989.86	998.78	991.49	23	1010	970.31	1001.04	1014.07	1003.85
11	996	970.31	990.72	999.96	992.44	24	987	970.31	1001.90	1015.24	1004.80
12	980	970.32	991.58	1001.13	993.39	25	1019	970.31	1002.76	1016.42	1005.75
13	982	970.31	992.44	1002.31	994.34	26	1013	970.31	1003.62	1017.60	1006.70

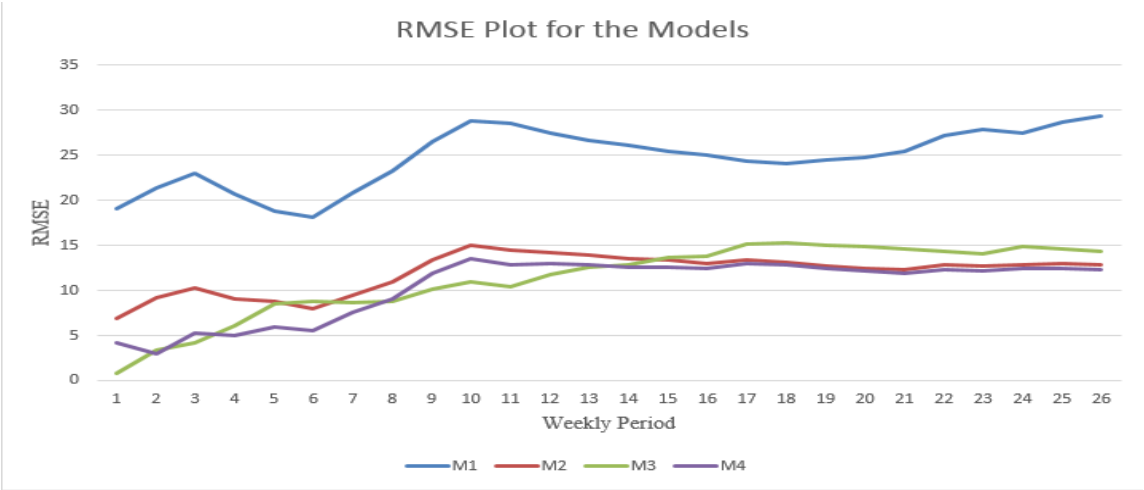
Table 4. MAPE, RMSE and MAD for the Four Models in Selected Intervals

	1 Week ahead forecast			12th Week ahead forecast			24th Week ahead forecast		
	MAPE	RMSE	MAD	MAPE	RMSE	MAD	MAPE	RMSE	MAD
M1	1.93	19.10	19.10	2.45	27.46	24.54	2.45	27.51	22.60
M2	0.70	6.88	6.88	1.20	14.21	12.03	1.09	12.82	10.01
M3	0.08	0.8	0.8	1.02	11.72	10.13	1.03	14.91	11.85
M4	0.42	4.14	4.14	1.04	12.93	10.40	1.02	12.42	9.43

Another interesting observation from figure 2 is that in the absence of M4, then M2 and M3 are the competing models. The two models show that M3 which was formulated by more past data values than M2 provides more accurate prediction in the first 14 weeks while M2 provides more accurate predictions in the weeks that follow. A conclusion drawn from this phenomenon is that in absence of M4, M3 will be a better model for short term forecasting and M2 will be a better model for a relatively long term forecasting.

Figure 3. Plot for MAPE and RMSE for the Four Models





5. Summary and Conclusion

The paper looked at the impact of the length of past observation in formulating ARIMA models and forecasting precision. In the formulated models, M1 is the worst model due to high levels of prediction errors. M2 predicted much better than M1 but it is however not the best model. Therefore to forecast one week ahead or 9 through 12 weeks ahead M3 may be used for better results. To forecast 2 weeks to 8 weeks and above 12 weeks, M4 may be used for better results. The conclusion from this study is that using too few past data values reduces the precision of forecasted values. However, according to the results of this study, increasing the length of past data in formulating the models does not necessarily produce the best results. A challenge to further research in this area is may be to come up with theoretical principles of identifying the length of historical data that will yield the best forecast for a given set of historical observations. It is therefore recommended that to obtain best ARIMA model, forecaster should consider both competing models as suggested by Box – Jenkins methodology and models formulated based on different data lengths always starting with the most recent data values.

6. References

Alba, E. and Mendoza, M. (2007). "Bayesian Methods for Forecasting Short Time Series". International Journal of Applied Forecasting. Issue 8: Fall 2007.

Al-Masudi, R. K.M., (2011): "Fitting ARIMA Models for Forecasting to inflow of Dokan Reservoir", Journal of Babylon University, Vol.19, No.4.

Biswas, R. and Bhattacharyya, B. (2013). "ARIMA Modeling to Forecast Area and Production of Rice in West Bengal". *Journal of Crop and Weed.* 9(2): 26 – 31 (2013)

Chatfield, C. (2000). Time – Series Forecasting. Boca Raton, Florida: Chapman and Hall/CRC

Ebenezer, O., Emmanuel, H., & Ebenezer, B. (2013). Forecasting and Planning for Solid Waste Generation in the Kumasi Metropolitan Area of Ghana. An ARIMA Time Series Approach. *International Journal of Sciences, Volume 2, Issue April 2013.*

Hanke, J.E. and Wichern, D.W. (2008). Business Forecasting, 8th edition. New Delhi: Pearson Education.

http://www.uobabylon.edu.iq/uobcoleges/fileshare/articles/Microsoft%20Word%20-%20Dokan(ARIMA).pdf

Khan, T. and Rahman, A. (2013). "Modeling the Contribution of Industry to Gross Domestic Product of Bangladesh". *International Journal of Economic Research.* Vol.4i2, 66 – 76.

Ljung, G. and G. E. P. Box. "On a Measure of Lack of Fit in Time Series Models." *Biometrika*. Vol. 66, 1978, pp. 67–72.

McLeod, A. I. and W. K. Li. "Diagnostic Checking ARMA Time Series Models Using Squared-Residual Autocorrelations." Journal of Time Series Analysis. Vol. 4, 1983, pp. 269–273.

Ofori, T., Ackah, B. and Ephraim, L. (2013). "Statistical Models for Forecasting Road Accident Injuries in

Ghana". *International Journal of Modern Mathematical Sciences*, 2013, 5(2):99 – 115.

Paul, J. C., Hoque, S. and Rahman, M. M. (2013). "Selection of Best ARIMA Model for Forecasting Average Daily

Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd". Global Journal of Management and Business Research Finance. Vol. 13, Issue 3, Version 1.0, 2013.

Sarpong, S. A. (2013). "Modeling and Forecasting Maternal Mortality: an Application of ARIMA models" *International Journal of Applied Science and Technology*. Vol. 3 No. 1; January 2013.

Singh, D. P., Kumar, P. and Prabakaran, K. (2013). "Application of ARIMA Model for Forecasting Paddy Production in Bastar division of Chhattisgarh". *American International Journal of Research in Science, Technology, Engineering and Mathematics.*

Wilberforce, K. (2012). Analysis of Gas Prices for Turkey From 2003 – 2011. Master's Thesis. Middle East Technical University.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar