Short Term Traffic Flow Forecasting by Time Series Analysis

A. HASNAT[1], F. I. RAHMAN[2]

[1]Department of Civil Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh (arif140092@gmail.com)
[2]Department of Civil and Environmental Engineering, University of Yamanashi, Kofu, Japan (ovi_faysal@yahoo.com)

Abstract

Time series analysis and forecasting has become a major tool in numerous applications. It has added a special feature of short-term prediction capacity in the history of forecasting. In this paper time series ARIMA (Autoregressive Integrated Moving Average) model is used to figure out the missing data points with traffic flow forecasting. The data used in this study are the traffic volume of Dublin Airport link road, Ireland of every 10 min interval of a day. This road is so important as 133773 vehicles in average use this road every day for airport purpose. Here using the traffic volume of previous 75% of a day, the least 25% traffic volume is forecasted. Then it is compared with the actual data of the forecasted hours. The ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) are plotted and checked the best model of fit for this data. From the analysis it has found that the required model is the Box-Jenkins ARIMA model which gives a better forecast. The mean absolute relative error (MARE) and mean absolute percentage prediction error (MAPPE) are calculated and it is 8.51% and 2.19% respectively. The R-squared value is 0.98. The four results are analyzed, summarized the variations with explanations and proposed the best model for the semi seasonal time series data.

Keywords: *time series analysis, flow, ARIMA, forecast, MAPE*

1 Introduction

Prediction science is generating it's importance day by day. It is now practiced in various fields for the convenience of human being's day to day task. Forecasting has made any operating any system much easier. In every stages of life like the closing stock price each day, birth rate at all hospitals in a city each year, product sales in units sold each day for a store, demand on a server each hour or number of passengers through a train station each day etc. all can be known previously using forecasting. Long term forecasts are used for system planning, scheduling construction of new generation capacity and purchasing of generating units (Jia et al., 2001). Intermediate term forecasts which is also called medium term forecast are used for maintenance scheduling, coordination of load dispatching and fixing of prices, so that demand can meet with original capacity (Jia et al., 2001). Adding with that, forecasting of next hours passengers of a station, number of approaching vehicles at next minute which is known as short term prediction is becoming popular because of the need of less data to forecast. Short-term forecasts are used for optimal generator unit commitment, fuel allocation, maintenance scheduling and buying and selling of power, economic scheduling of generating capacity, scheduling of fuel purchases, security analysis and short-term maintenance scheduling (Jia *et al*., 2001; Pedregal *et al*., 2010). Very short term forecast are used for security assessment and economic dispatching, real time control and real time security evaluation (Jia *et al*., 2001). The four main categories of time horizons have been studied extensively. In case of long term forecasts Al-Saba *et al*.(1999), Kermanshashi (2002) and Carpinteiro *et al*. (2007) used Artificial Neural network (ANN); intermediate term forecasts by Elkateb *et al*. (1998), Mirasgedis *et al*. (2006) and Tsekouras *et al*. (2007). They also used Artificial/Fuzzy Neural Network for forecasting; short-term forecasts by Al-Hamadi and Soliman (2004) based on Kalman filtering algorithm, Hobbs *et al*. (1998) and Catalao *et al*. (2007) based on neural network approach and Faysal Ibna Rahman. (2018) on Monte Carlo simulation; very short term forecasts by Taylor (2008) and Taylor *et al*. (2008).

Several forecasting methods including multiple linear regression Al-Hamadi (2005), Amjady and Keynia (2011), K. Prabakaran *et al*. (2013) and Mirasgedis *et al*, (2006); nonlinear multivariable regression model Al Rashidi and

El-Naggar (2010), Tsekouras *et al*. (2007) and Suwardo *et al*. (2010) are implemented for different types of forecasting and varying degrees of success. Besides the time series methods AR(Autoregressive), MA(Moving Average), ARMA(Autoregressive Moving Average) and ARIMA(Autoregressive Integrated Moving Average) etc. have proved it's capacity of standard forecasting in recent years.

In case of the practical need of knowing the number of vehicles approaching in different hours of a busy road, short term traffic flow has gained great attention compared to others. Because the accuracy of prediction affects the maintaining capacity of traffic operation that results in a great traffic jam. Since maintaining the appropriate load of traffic is crucial for the city it is extremely essential to forecast an accurate traffic flow. That's why time series ARIMA model has been implemented here. Since different time series models have different characteristics with different types of data sets (Hamilton, 1994), here it has been tried to find out the possibility and limitations of ARIMA model to fit this type of data.

2 Methodology

2.1 Study location and data collection The data worked out here have been found from the Transport Infrastructure Ireland (TII). The traffic volume used here is from the link road between junction-1 and junction2 of Dublin Airport route, Ireland from January 31, 2018 to February 03, 2018 [Fig.1].

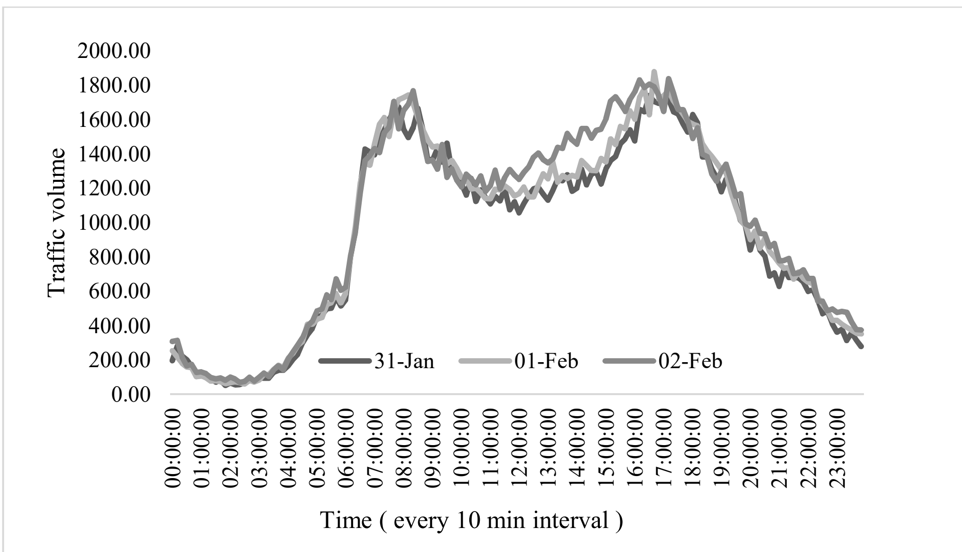


Figure 1. Observed original traffic volume with respect to time

In this paper, Box-Jenkins ARIMA model has been implemented to forecast traffic volume. Here the type of forecast accomplished is using 75% data of 31 January and forecasted the least 25% of the same day. And then the forecasted value is compared to the original value of 31 January. Another type of forecasting can be done through same procedure which is the prediction of 01 February whole day volume using the total volume of 31 January. Here also the forecasted value should be compared with the real one.

To fit the non-seasonal Box-Jenkins ARIMA model for a stationary time series data there are some steps. The forecasting follows directly from the fitted model. The general form of ARIMA (p,d,q) model can be represented as: $Z_t = Ø_1 Z_{t-1} + Ø_2 Z_{t-2} + \ldots\ldots + Ø_p Z_{t-p} - \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$;

Where, Z_t is the value of a stationary time series at time t and ε_t's represent random error which are being independently and normally distributed with zero mean and constant variance for time t = 1, 2…..n. 'p' is for autoregressive order; 'd' is number of differencing and 'q' is the moving average order; Ø's and θ's are co-efficients to be estimated.

i) Model identification

The order of the model is identified based on time domain and frequency domain analysis i.e. autocorrelation function (ACF), partial autocorrelation function (PACF) and spectral density function. A graph of autocorrelation function determines whether the series is stationary or not. The time series is considered stationary if the graph of ACF values either cuts off fairly quickly or dies down fairly quickly. The series is considered as non-stationary if the graph of ACF dies down extremely slowly. In case of the non-stationary series, it can be converted to a stationary series by successive differencing (Fig.2). Besides the stationarity can also be checked by trend analysis through mean and variance [Table1]. In this case the 1st difference of log(flow) gives almost zero mean and 1st difference of log(flow) gives most smaller variance. So, ACF and PACF graph are plotted considering d(log(flow)) of data[Fig.3]

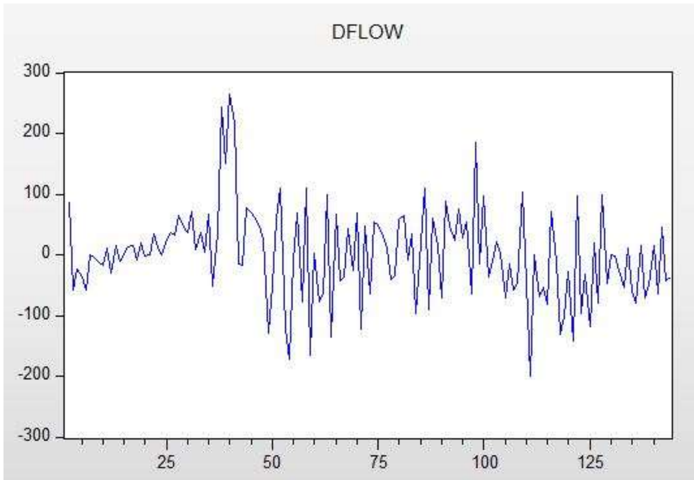
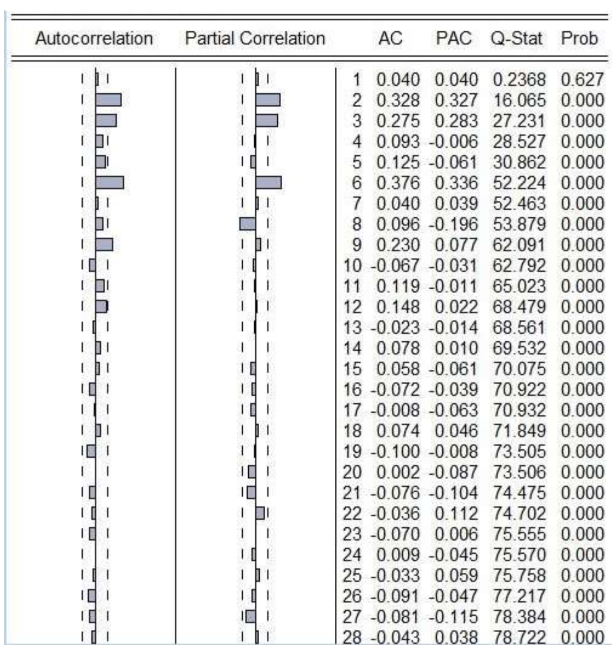


Figure 2. Graph of the data after 1st difference (Stationarity check)

After 1st difference, the mean of the data turns to almost zero and the variance is nearly constant [Table 01] and trend is almost invisible [Fig.2]. So, there is no need of more difference. Thus the difference order becomes '1'.

Table 01: Stationarity analysis of data

	Mean	**Variance**
flow	928.97	289130.03
d(flow)	0.5734	5638.7201
log(flow)	6.5082	0.9622505
d(log(flow))	0.0024	0.0144999



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.040	0.040	0.2368	0.627
		2	0.328	0.327	16.065	0.000
		3	0.275	0.283	27.231	0.000
		4	0.093	-0.006	28.527	0.000
		5	0.125	-0.061	30.862	0.000
		6	0.376	0.336	52.224	0.000
		7	0.040	0.039	52.463	0.000
		8	0.096	-0.196	53.879	0.000
		9	0.230	0.077	62.091	0.000
		10	-0.067	-0.031	62.792	0.000
		11	0.119	-0.011	65.023	0.000
		12	0.148	0.022	68.479	0.000
		13	-0.023	-0.014	68.561	0.000
		14	0.078	0.010	69.532	0.000
		15	0.058	-0.061	70.075	0.000
		16	-0.072	-0.039	70.922	0.000
		17	-0.008	-0.063	70.932	0.000
		18	0.074	0.046	71.849	0.000
		19	-0.100	-0.008	73.505	0.000
		20	0.002	-0.087	73.506	0.000
		21	-0.076	-0.104	74.475	0.000
		22	-0.036	0.112	74.702	0.000
		23	-0.070	0.006	75.555	0.000
		24	0.009	-0.045	75.570	0.000
		25	-0.033	0.059	75.758	0.000
		26	-0.091	-0.047	77.217	0.000
		27	-0.081	-0.115	78.384	0.000
		28	-0.043	0.038	78.722	0.000

Figure 3. Graph of ACF and PACF using EVIEWS software

Therefore, based upon the conditions of values and graphical plot of ACF and PACF, it follows the following tentative ARIMA(p,d,q) models shown in table 2. To select the best suitable model for forecasting out of the models proposed, the lowest BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) values are needed. It is found that the ARIMA(6,1,0) shows the smallest AIC and BIC value among other proposed models [Table 2]. Thus ARIMA(6,1,0) is considered as the best model of fit for this data and the forecasting is carried out using this fit.

Table 2. AIC and BIC values of fitted ARIMA model

		ARIMA Model						
		(2,1,0)	**(0,1,2)**	**(3,1,0)**	**(0,1,3)**	**(6,1,0)**	**(2,1,2)**	**(0,1,6)**
Co efficients	AR(1)	0.08		-0.04		-0.08	0.08	
	AR(2)	0.32		0.36		0.31	0.32	
	AR(3)			0.24		0.20		
	AR(4)					-0.006		
	AR(5)					-0.08		
	AR(6)					0.31		
	MA(1)		-0.09		-0.01		0.098	0.036
	MA(2)		0.56		0.53		0.566	0.414
	MA(3)				0.13			0.069
	MA(4)							0.032
	MA(5)							0.093
	MA(6)							0.563
	Log-Likelihood	100.80	101.75	106.10	102.37	121.78	101.275	111.37
	AIC	-1.55	-1.54	-1.63	-1.53	**-1.88**	-1.54	-1.63
	BIC	-1.48	-1.47	-1.54	-1.44	**-1.72**	-1.47	-1.47

ii) Model estimation

ARIMA fitting order p, d, q values and their statistical significance can be judged by t-distribution. A model with minimum values of RMSE, MAPE, AIC, BIC, Q-statistics and with high R-square, may be considered as an appropriate model for forecasting. The model selection criteria includes Akaike Information criterion (AIC) and Schwarz's Bayesian Information criterion (BIC), Mean squared error (MSE), Root Mean squared error (RMSE), Mean absolute error (MAE) and Minimum Absolute Percentage Error (MAPE)

iii) Diagnostic checking

It is necessary to ensure the residuals estimated from the model are white noise. So the autocorrelations of the residuals are to be estimated for the diagnostic checking of the model. These may also be judged by Ljung-Box statistic under null hypothesis that autocorrelation co-efficient is equal to zero. Moreover, it can also be checked that the properties of the residual with the graph as follows.

1) Check the normality by considering the normal probability plot or the p-value from the One-Sample Kolmogorov – Smirnov Test.

2) Check the randomness of the residuals by considering the graph of ACF and PACF of the residual. The individual residual autocorrelation should be small and generally within ± 1.96/ N of zero.

iv) Forecast

ARIMA models are developed basically to forecast the corresponding variable. The entire data is segregated in two parts, one for sample period forecasts and the other for post-sample period forecasts. Frequency domain analysis is one of the first analytical techniques developed by Koreisha and Fung (1999) and Pankratz (1983). It is also known as periodogram analysis. Evaluation of forecasting found from the periodogram analysis was performed by using mean absolute relative error (MARE) and mean absolute percentage prediction error (MAPPE).

3 Result and Discussion

Dublin airport link road is one of the busiest roads of Ireland. This road is so important as 133773 vehicles in average use this road every day. So, number of vehicle moving at every minute is important for traffic operation for the authority and knowing the volume of traffic of upcoming hours play an vital role to take effective measures. That's why using the data of 75% time of the day, the least 25% is predicted here. The prediction scenario is shown in [Fig.4]

Here using the data of 00:00:00 to 21:10:00, the least 21:20:00 to 23:40:00 are predicted using Box-Jenkins ARIMA(6,1,0) model and the forecasted data is compared with original data to measure the errors [Table 3] The performance of model is measured by the degree of accuracy. Accuracy of the model is indicated by statistical closeness such as mean absolute relative error (MARE) and mean absolute percentage predicting error (MAPPE). Both are indicator of model performance. The model which has minimum value of MARE and MAPPE is the accurate model (the best) among the several tentative models in predicting. In other word, the minimum residual (error) indicate high accuracy model.

Table 03: Forecasting and Forecasted Error calculation

Time	Original data	Forecasted	MARE	MAPPE
21:20:00	681	706.4292	1.69528	0.24894
21:30:00	681	678.7646	0.149025	0.02188
21:40:00	679	697.1039	1.206925	0.17775
21:50:00	654	673.8079	1.320525	0.20191
22:00:00	600	698.0589	6.537262	1.08954
22:10:00	610	675.4425	4.3628	0.71521
22:20:00	549	675.7625	8.450832	1.53931
22:30:00	470	595.6278	8.375131	1.78194
22:40:00	484	563.9558	5.330384	1.10131
22:50:00	413	606.7447	12.91652	3.12748
23:00:00	363	534.8869	11.45913	3.15684
23:10:00	379	529.8712	10.05808	2.65384
23:20:00	314	603.0022	19.26681	6.13592
23:30:00	369	607.9601	16.59734	4.62321
23:40:00	316	617.0552	20.07035	6.35137

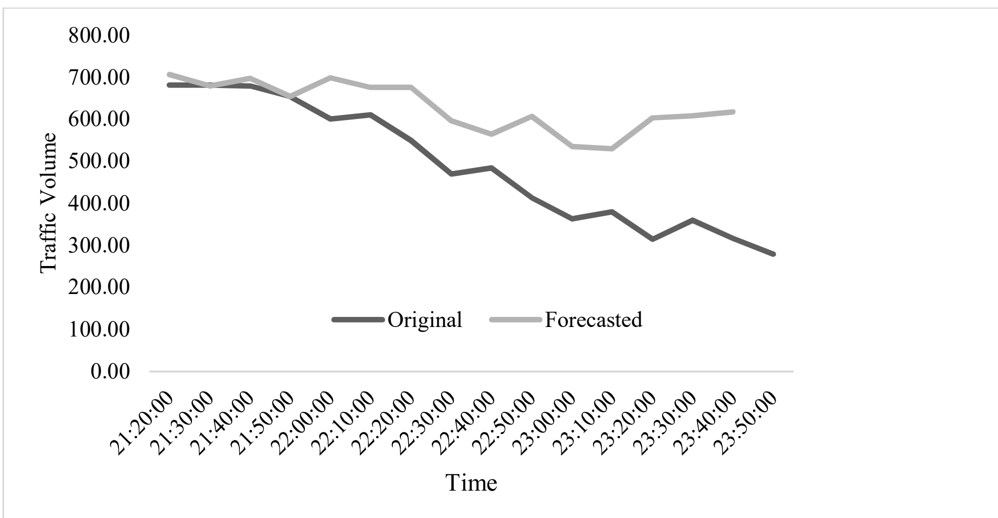


Figure 4. Original vs forecasted graph using ARIMA model.

4 Conclusion

The data is semi seasonal and non-stationary. Operating 1st difference makes the data stationary and partially meets the requirements of fitting Autoregressive Integrated Moving Average model. The results show that the forecasting of traffic volume using time series ARIMA model is nearly same with the original volume. The residuals prove that it can be implemented in practical field. But as the data has some seasonality and ARIMA method is for non-seasonal data sets that's why a little regression occurs here which can be understood from the forecast of 22:00:00 to 23:40:00. Adding with that time series can't forecast abnormal change in data points whereas there is a steep gradient contained in 23:10:00 to 23:40:00 of original data. This deviation results in a steeper change in that prediction region. So, implementing a seasonal ARIMA model may overcome this horizon.

5 References

Al-Saba, T., and El-Amin, I. (1999). Artificial neural networks as applied to long-term demand forecasting. Artificial Intelligence in Engineering, 13(2), 189–197. doi:10.1016/s0954-1810(98)00018-1

Al-Hamadi, H. M., and Soliman, S. A. (2005). Long-term/mid-term electric load forecasting based on short-term correlation and annual growth. Electric Power Systems Research, 74(3), 353–361. doi:10.1016/j.epsr.2004.10.015

Amjady, N., & Keynia, F. (2011). A new prediction strategy for price spike forecasting of day-ahead electricity markets. Applied Soft Computing, 11(6), 4246–4256

Box, G.E.P., and Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, California, 1970

Fausto P. Garcia, Diego J. Pedregal (2010). Time series methods applied to failure prediction and detection reliability engineering & system safety. Elsevier, volume 95, Issue 6, June 2010, Pages 698-703.

Faysal Ibna Rahman. (2018). Short term traffic flow prediction by Monte Carlo simulation. Transport & Logistics: the International Journal, 2018. Volume 18, Issue 44, July 2018, ISSN 2406-1069.

Kermanshashi and Iwamiya (2002). Analysis of Least Square and Exponential Regression Techniques for Energy Demand Requirement (2013-2032). American Journal of Electrical and Electronic Engineering. 2018; 6(2):38-59. doi: 10.12691/ajeee-6-2-1.

Kumar Manoj. (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. Studies in Business and Economics.

K. Prabakaran *et al*. (2013). Forecasting Cultivated Areas And Production Of Wheat In India Using ARIMA Model. Golden Research Thoughts. ISSN 2231-5063. Volume-3, Issue-3, Sept-2013.

Mirasgedis *et al*. (2006). Models for mid-term electricity demand forecasting incorporating weather influences. Energy 31, 208–227.

R Biswas and B. Bhattacharyya (2013).ARIMA modeling to forecast area and production of rice in West Bengal. Journal of Crop and Weed, 9(2):26-31(2013)

Suwardo *et al*. (2010). ARIMA models for bus travel time prediction. Journal-The Institution of Engineers Malaysia. Vol. 71, No.2, June 2010.

Sigauke, C., & Chikobvu, D. (2011). Prediction of daily peak electricity demand in South Africa using volatility forecasting models. Energy Economics, 33(5), 882–888. doi:10.1016/j.eneco.2011.02.013

Taylor, J.W., 2008. An evaluation of methods for very short-term load forecasting using minute-by-minute British data. International Journal of Forecasting 24, 645–658.

Zhang, J., Tan, Z., & Yang, S. (2012). Day-ahead electricity price forecasting by a new hybrid method. Computers & Industrial Engineering, 63(3), 695–701. doi:10.1016/j.cie.2012.03.016