Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 48 (2020) 3880–3895

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Exploratory Study on Approaches for Traffic count Prediction; Using Toll-Way Traffic Count

Soorya V. B.[a] , Shriniwas S. Arkatkar[b,*], Lelitha Vanajakshi[c]

[a] *Masters student, Transportation Engineering and Planning Section, Civil Engineering Department, SVNIT-Surat, Gujrat-395007, India*
[b] *Assistant Professor, Transportation Engineering and Planning Section, Civil Engineering Department, SVNIT-Surat, Gujrat-395007, India*
[c] *Professor, Department of Civil Engineering, Indian Institute of Technology Madras,Chennai 600 036, India*

Abstract

The process of predicting or simulating traffic conditions, based on current and past traffic observations is an important component of any of the Intelligent Transportation System (ITS) applications. There are several methods to predict traffic count. However, each method in different literatures may use different datasets, different time intervals of input traffic flow/count. One of the aims of this study is to provide a review and performance analysis of parametric and non-parametric approaches on traffic prediction. Second, to explore the possibilities in the implementation of Advanced Traffic Management Systems (ATMS), one of the functional areas of Intelligent Transportation Systems (ITS), by predicting traffic count on toll plazas for optimizing of toll-plaza operations. An RFID (Radio-Frequency Identification) based Electronic Toll Collection (ETC) system gives timely varying traffic counts observed at toll plazas, which has been utilized to develop prediction models based on historic data. An empirical differentiation of four methods, namely Seasonal Autoregressive Integrated Moving Average (SARIMA) model based on time series analysis, Monte Carlo Simulation (MCS), Random Forest (RF) based on tree ensemble learning technique and KNN non-parametric regression-based machine learning technique, are proposed. Performance analysis at varying time intervals (5, 10, and 15 minutes) of input traffic count, for all the aforesaid models were compared with Simple Average Technique (SAT) using the historic data collected from two different toll-plazas in India. It was observed that K Nearest Neighbors (KNN) non-parametric regression performed better than other methods in most of the cases.

Keywords: Tolling application; Traffic volume prediction; Seasonal Autoregressive Integrated Moving Average (SARIMA); Monte Carlo Simulation (MCS); Random Forest (RF) tree ensemble learning technique; KNN non-parametric regression;

* Corresponding author. Tel.: +91 8140252777;
E-mail address: sarkatkar@ced.svnit.ac.in

10.1016/j.trpro.2020.08.033

1. INTRODUCTION AND BACKGROUND

Intelligent Transportation Systems (ITS) enhance the efficacy and safety of transportation systems using the modern technologies. The ability to accurately predict current traffic conditions, which is a critical component of Advanced Traffic Management Systems (ATMS) applications, is a fundamental part of ITS (Ben-Akiva et al. (1998)). The provision of timely and accurate traffic information is valuable for both operators and users of the infrastructure. The installation of traffic management centres that imparts this information would improve the efficiency, performance and coordination of all traffic operations. However, successful implementation of such advanced systems requires a thorough understanding of historical, present and future traffic conditions of the system. In order to control the transportation system in a proactive manner, there will not be any lag between the collection of data and implementation of traffic control strategies (Smith et al. (2002)). Development of models and algorithms which are suitable for implementation of ITS is an important task for the researchers in this area. Aptly, the importance of short-term traffic prediction module in ATMS architecture was highlighted by Barcelo and Garcia (1991). Several techniques were available for prediction, testing of the practical accuracy is very important. Aim of this study is two-fold, to review performance analysis of popular prediction techniques using toll way traffic data, based on these comparisons to propose a best toll way traffic count prediction system. Second, to explore the possibilities in the implementation of Advanced Traffic Management Systems (ATMS), one of the functional areas of ITS, by predicting traffic count on toll plazas for optimization of toll-plaza operations. Generally, the traffic count has been found to vary with time at toll plazas (Chao and Xiuli (2000)). Effective toll plaza operation is very important during both the peak and the off-peak hours. Suboptimal operation during peak hours adversely affects the throughput of the facilities, whereas, during hours with low traffic, it may result in high operating costs.

Electronic Toll Collection (ETC) is one of the major user services under ITS and aims for easy access of vehicles in peak hours and concurrently reducing the operating cost in off-peak hours. The main functional components of the Electronic Toll Collection (ETC) are Automatic Vehicle Identification (AVI), Automatic Vehicle Classification (AVC) and Vehicle Enforcement System (VES) (Muthulakshmi (2015)). For this, the type of vehicles, date and time, lane number etc. are recorded. Analysis of this data and prediction of traffic volume to future times are the main objectives of the present study. Prediction of traffic count using different parametric and non-parametric approaches such as Seasonal Autoregressive Integrated Moving Average (SARIMA) model based on time series analysis, Monte Carlo Simulation (MCS), Random Forest (RF) tree ensemble learning technique, and K Nearest Neighbors (KNN) non-parametric regression are tried out. An effort has been made to compare the performance of all these models using the same set of data and same interval of input traffic count. Data from two toll locations were used and vehicular flow, aggregated into 5, 10, and 15-minute intervals, has been used as input. Also, by inputting total daily traffic count of previous months, next month daily traffic count has been predicted. Which can be used to estimate monthly revenue at toll plazas in advance. The performance of the aforesaid approaches was compared to a baseline Simple Average Techniques (SAT).

At the early stages of research in this topic, the methodology was constrained by the availability of reliable and continuous data. With the development of technology, data collection has become less tedious and a large amount of data is available for analysis. Many researches have been reported in this regard to develop accurate prediction models using the data, most of which are developed for homogeneous traffic conditions.

Application of time series analysis for traffic forecasting was reported by many. Some of the initial studies carried out by Ta-Yin Hu et al. (2010), Ghosh et al. (2009) and Wang et al. (2011), indicates the improved prediction accuracy by using time series analysis for short term traffic count prediction. Jha et al. (2016) developed ARIMA models for traffic forecasting. Kumar and Vanajakshi (2015) developed SARIMA model for short term traffic flow prediction, using historic traffic flow data.

Raychaudhuri (2008) and Cheah et al. (2006), conveyed applications of MCS in various engineering disciplines for forecasting. Mishraa et al. (2015) applied MCS for short term traffic count prediction. Other prediction techniques reported include machine learning techniques and non-parametric regression. Breiman (2001) introduced an effective tool in prediction called RF machine learning technique. Leshem et.al (2007) predicted traffic flow by combining Random Forests algorithm into Adaboost algorithm. Hamner et.al (2010) used Random Forests by modeling local and aggregate traffic flow. Smith (2002) examined theoretical foundation of non-parametric regression and compared with Seasonal ARIMA models. They found a significant improvement in performance of the non-parametric regression (K-

NN). Wang et.al (2015) used a hybrid model of K-NN and SVM (Support Vector Machine) for short-term prediction of freeway exiting traffic flow.

There are several methods to predict traffic count. However, each method in different literatures may use different datasets, different time intervals of input traffic flow/count. Also, it can be seen that majority of these studies were from homogeneous and lane-based traffic conditions. In developing countries like India, the traffic is composed of different categories of vehicles, making it highly heterogeneous in nature and difficult to account for. An exploratory comparison of aforesaid methods to predict traffic flow has been carried out in the present study. In addition, traffic count prediction at toll plazas have not been explored in detail. Based on these gaps, the present study aims to predict the traffic count at toll plazas, under the heterogeneous traffic conditions. BRIANL (2003) suggested that with decreasing of the forecasting time interval, the prediction accuracy would become worse. The present study examined the effect of different forecasting time interval in prediction accuracy of each model. The study uses time series analysis (SARIMA model), Monte Carlo simulation, random forest technique, and KNN regression. A brief theoretical description of aforesaid prediction techniques, is presented in the following section.

2. PREDICTION TECHNIQUES

There are generally two kinds of techniques for short term forecasting- parametric and non-parametric approaches. The parametric approaches develop direct forecasting mathematically by a set of parameters. Parametric model fitting effectively compresses all data in the training set (historical data) into one equation and then find out a group of parameters which can minimize the forecasting error. Afterwards, the model can be used for forecasting. Nonparametric regression, on the other hand, is a data-driven approach, which retains all historical observations and searches for the most similar case of the current state, based on which forecasting is made. For the present study, to predict traffic count on toll plazas, the following techniques are used.

1. SARIMA model, based on time series analysis
2. Monte Carlo Simulation
3. Random Forest (RF) based on tree ensemble learning technique
4. KNN regression-based machine learning technique

2.1 Time series analysis

Time series is a sequence of measurements over time. Statistical properties such as mean, variance, auto covariance etc. are mainly considered in these analyses. There are different types of time series techniques. Three basic models are Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA) (Wang et al. (2015)). Stationarity of the series is one basic assumption made in many time series analysis techniques. For non-stationary series, regular differencing is applied to convert stationary series together with AR and MA and is referred to as Autoregressive Integrated Moving Average (ARIMA) model. If the series contains a strong seasonal pattern with periodic fluctuation, differencing at the lag specified by the seasonal period is required in order to convert the series to a stationary series and such a series is modelled by Seasonal Autoregressive Integrated Moving Average (SARIMA) models. The SARIMA model is represented by ARIMA $(p, d, q) \times (P, D, Q)\, S$, where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order and, S is the time span of repeating seasonal pattern. For SARIMA models, in order to convert the input time series data to a stationary process, usually differencing at the lag specified by the seasonal period is sufficient. If the data contains trend and seasonality, then differencing at lag-1 (if trend is linear) as well as the differencing at the lag specified by the seasonal period is required. The general representation of SARIMA model is given by

$$\emptyset_{\mathrm{p}}(B^s)\emptyset(\mathrm{B})\nabla_s^D\nabla^d x_t = \delta + \theta_Q(B^s)\theta(B)\omega_t, \quad (1)$$

where, ω_t is the Gaussian white noise process, polynomials $\emptyset$(B) and θ(B) of orders p and q, respectively are the ordinary autoregressive and moving average components, $\emptyset_{\mathrm{p}}(B^s)$ and $\theta_Q(B^s)$ of orders P and Q are the seasonal

autoregressive and moving average components and $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^s)^D$ is the ordinary and seasonal difference components.

Model identification, model estimation, diagnostic checking, and forecasting and validation are the important steps involved in time series modeling. The first step in model identification is to determine whether the series is stationary and, if there is any significant seasonality that needs to be modelled. There are mainly two ways to check stationarity of a given series; using plot of Auto Correlation Function (ACF) and Dicky-Fuller unit root test (Shumway and Stoffer (2000)). The diagnostic usually uses the Akaike Information Criterion (AIC), and the maximum likelihood estimation (MLE). The 'auto.arima' function in R software, which combines unit root tests, minimization of AIC and MLE to obtain an ARIMA model, was employed. For the diagnostics of the fitted model the residuals of the chosen model are checked by plotting the ACF of the residuals. Once, the residuals look like white noise, forecasts can be made.

2.2 Monte Carlo Simulation (MCS)

It is a type of simulation that depends on repeated random sampling and statistical analysis to compute the results. Input distribution identification (distribution fitting) and random number generation are the important steps in this process (Wang and Chen (2011)). In MCS, the first step is to identify the best fitting statistical distribution for the random variables, which can be done using Maximum Likelihood Method. Each probability distribution can be uniquely identified by its parameter set. Next step is to generate random numbers between 0 and 1. The prediction can be done using the random numbers generated and the parameters of the fitted statistical distributions. For each set of input values, one set of output value can be generated. By collecting output values from several simulations runs, we can perform statistical analysis on the values of the output parameters.

2.3 Random Forest (RF) based on Tree Ensemble learning technique

RF develops many decision trees based on random selection of data and variables during training. The class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Breiman (2001)) will be the corresponding output. A decision tree creates a type of flowchart which consists of nodes (referred to as 'leafs') and a set of decisions to be made based on these node (referred to as 'branches'), where each split corresponds to a node in it. It is an extensive modification of bagging (bootstrap aggregation) that builds a large collection of de-correlated trees, and then averages it.

The bootstrap is a powerful statistical method for estimating a quantity from a data sample that relies on random sampling with replacement. Bagging or bootstrap aggregation is a technique for reducing the variance of an estimated prediction function. In bagging process, algorithm generates a number of new training data sets. In this study, the regression model developed based on Random forest ensemble learning method are built using the 'randomForest' class in R software. Fig. 1 shows the flowchart of Random forest ensemble learning technique.

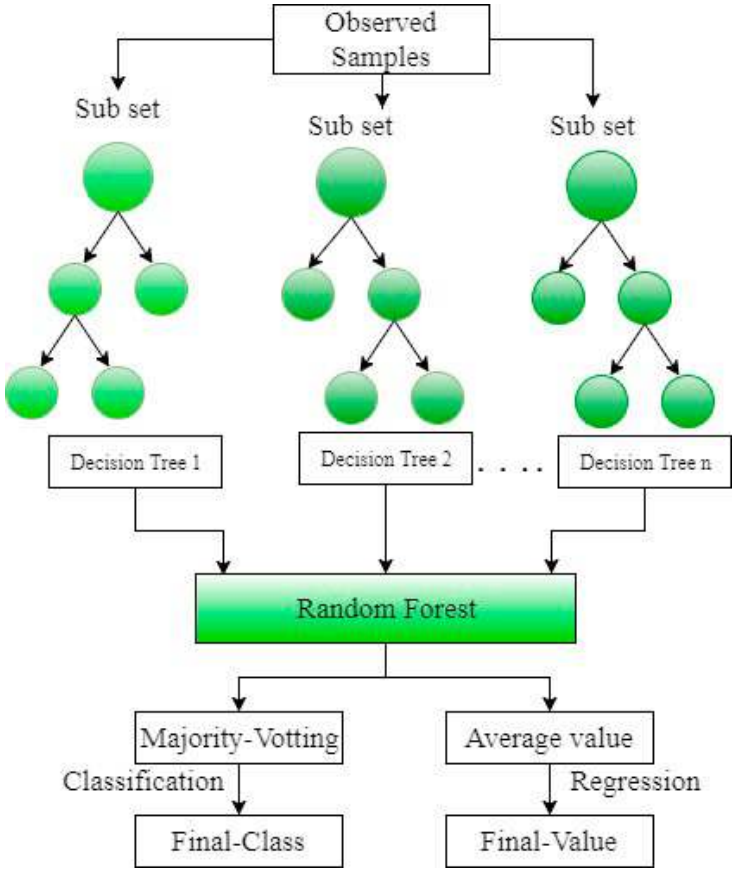

Fig. 1 Random Forest Ensemble learning technique

2.4 K-nearest neighbor (KNN) Technique

K-NN is a non-parametric regression-based forecasting technique, which searches a collection of historical observations for records like the current conditions and uses these to estimate the future state of the system. It has three components, such as historic database, search procedure and forecast generation. The historical database is a collection of past observations, which in the present study were the traffic counts collected from tollways. For univariate time series data, a system can be defined with a measurement at time t, t-1, …, t-D, where D is the number of lags. For example, X(t) of measurements collected every ten minutes with D = 2 can be written as

$$X(t)=[V(t),V(t-1),V(t-2)] \quad , \tag{2}$$

where V(t) is the measurement during the current time interval, V(t-1) is the measurement during the previous interval, and so on.

The search procedure finds records in the historical database that are like the current conditions and labels them as neighbors. The distance between the stored data and the new instance is calculated using similarity measures (R. Keith et al. (2000)). For the present study, Euclidean distance was considered. Similarity measure from historical record 'p' to the current condition 'q' can be written as

$$L_2 = \left[\sum_{i=1}^{k} |p_i - q_i|^2\right]^{1/2} , \tag{3}$$

where, p_i is the i^{th} element of the historical record currently under consideration, and q_i is the i^{th} element of the current record. The search procedure finds the nearest neighbors, which are those historical observations with the smallest values of distance (p, q).

The final stage is forecast generation. The most commonly used approach for that is to compute a simple average of the dependent variable values of the nearest neighbors. Mathematically, this estimate is calculated according to Equation (4), where k is the number of neighbors found by the search procedure.

$$\hat{V}(t+1) = \left(\frac{1}{k}\right)\sum_{i=1}^{k} V_i\,(t+1)\,. \tag{4}$$

In this study, traffic count prediction using KNN algorithm was done using R software along with finding the optimum value of 'k' (number of nearest neighbors to be considered for averaging). The Fast-Nearest Neighbor Search Algorithms and Applications (FNN) package available in R software was made use of.

3. DATA COLLECTION AND ANALYSIS

The data collected using the electronic entry system at two different toll plazas in India, (i) Sanand toll plaza, which is located on NH 147, Sarkhej-Viramgam section in the outskirts of the city of Ahmedabad in Gujrat, and (ii) Krishnagiri - Thoppur toll plaza, which is located on NH-44, Krishnagiri Hosur highway in Tamil Nadu, India, were used in this study. The data collected from Sanand toll plaza is from 6th October 2015 to 13th October 2015.The raw data from the electronic entry system of the toll plaza contained every five-minute class-wise traffic flow for the entire 24 hours from 12 midnight to 12 midnight. From the second location, similar data was obtained from 1st April 2017 to 30th April 2017. The total number of vehicles aggregated into five-minute, ten-minute, and fifteen-minute time intervals, were considered as input. The prediction models have been validated using both data sets, one day data was kept for validation of the models from collected data for the input time intervals mentioned above. Since the second location have one-month data the practical application of these techniques has been tested using this data. Also, another set of total daily traffic count from second location was collected from 1st April 2015 to 31st March 2016, to predict daily traffic count. According to the availability of the data the prediction models were developed and validated for each location.

Traffic count is expected to follow certain temporal patterns, and hence, traffic count pattern analysis was conducted first. Temporal patterns can include hourly pattern (peak and off-peak), daily pattern (week day vs week end), weekly pattern (same days of the week having similar pattern) etc., Identification of these patterns in the data will help in identifying the best inputs to be used for the prediction application. Hourly pattern can be explained as the variation in traffic count over the hours of a day, mainly to capture the peak and off-peak variations. In daily patterns, difference between weekday and weekends can be checked and in weekly pattern, the variation in traffic count for the same day of the week was analyzed. Box plots were made to understand the traffic count variation on hourly, daily, weekly basis, using R-software and are shown in Fig. 2 and 3 for two locations, selected in the present study.

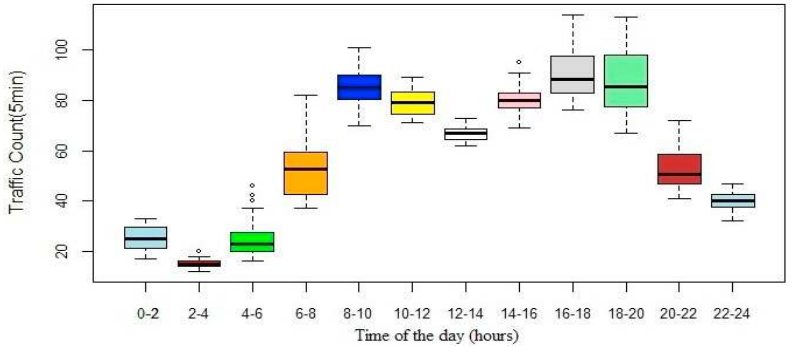

Fig. 2 Hourly box plot of traffic counts for Sanand toll plaza (Location 1)

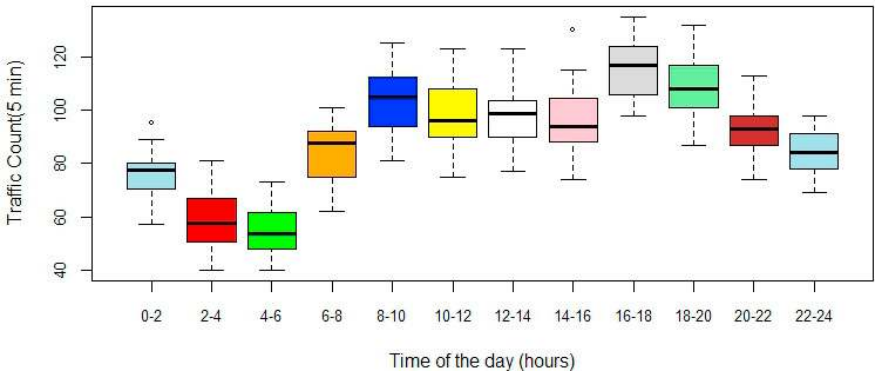

Fig. 3 Hourly box plot of traffic counts for Krishnagiri-thopur toll plaza (Location 2)

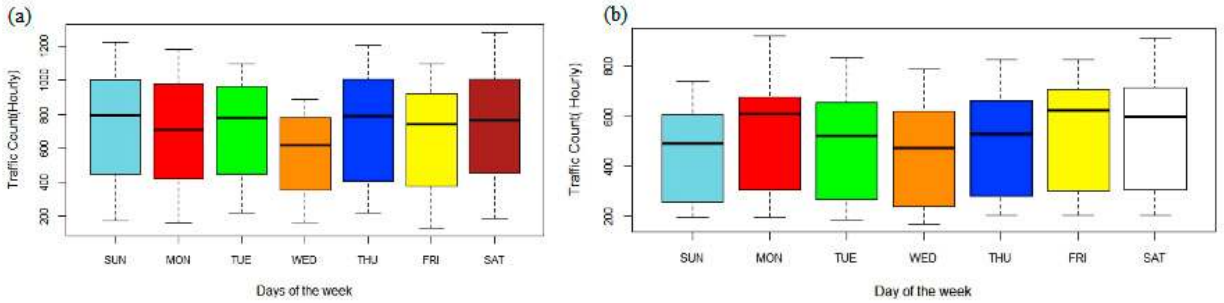

Fig. 4 Daily box plot of traffic counts for Krishnagiri-thopur toll plaza ((a) Location 1 &(b) Location 2)

From Fig. 2 and Fig. 3, it can be seen that there is a distinct variation in the traffic counts for each two-hour interval over the day. There is a morning peak in traffic count during 8-10 hours and during 16-20 hours in the evening. A prediction model for hourly traffic count for a day is expected to capture these variations. Similar analysis for daily variation was also made and a significant variation was observed. Figure 4 shows daily box plot for the variation of traffic count over the days of the week. Different time intervals of input data for all the models have been tried, and one sample prediction model for each method is explained as follows and all other set results are given in results and discussion part.

4. PREDICTION MODELS

4.1 SARIMA model

To capture the daily pattern of traffic, the 24 hours' data collected from Location 1, for the 6th, 7th, and 8th of October 2015 was used for model development and the developed model was validated using the observed flow data of October 9, 2015. Since, one of the studies concluded that previous three days flow as input is adequate for predicting the next 24 hours ahead flow (Kumar and Vanajakshi, (2015)).

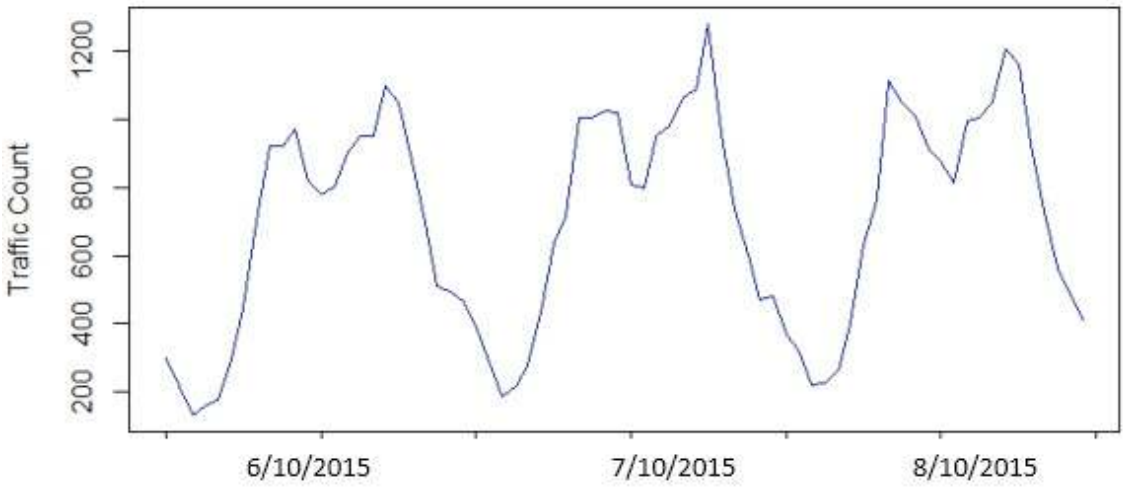

Fig. 5 Time series plot of three-day traffic count at Location 1

The first step in model development was to plot the time series data and check for any patterns. The time series plot of the observed hourly vehicle count of these three consecutive days, is shown in Fig. 5, in which there is a clear seasonal pattern with seasonality of 24 hours. Also, ACF and PACF of the data were plotted (Fig. 6 (a) and (b)) to study the seasonality and stationarity.

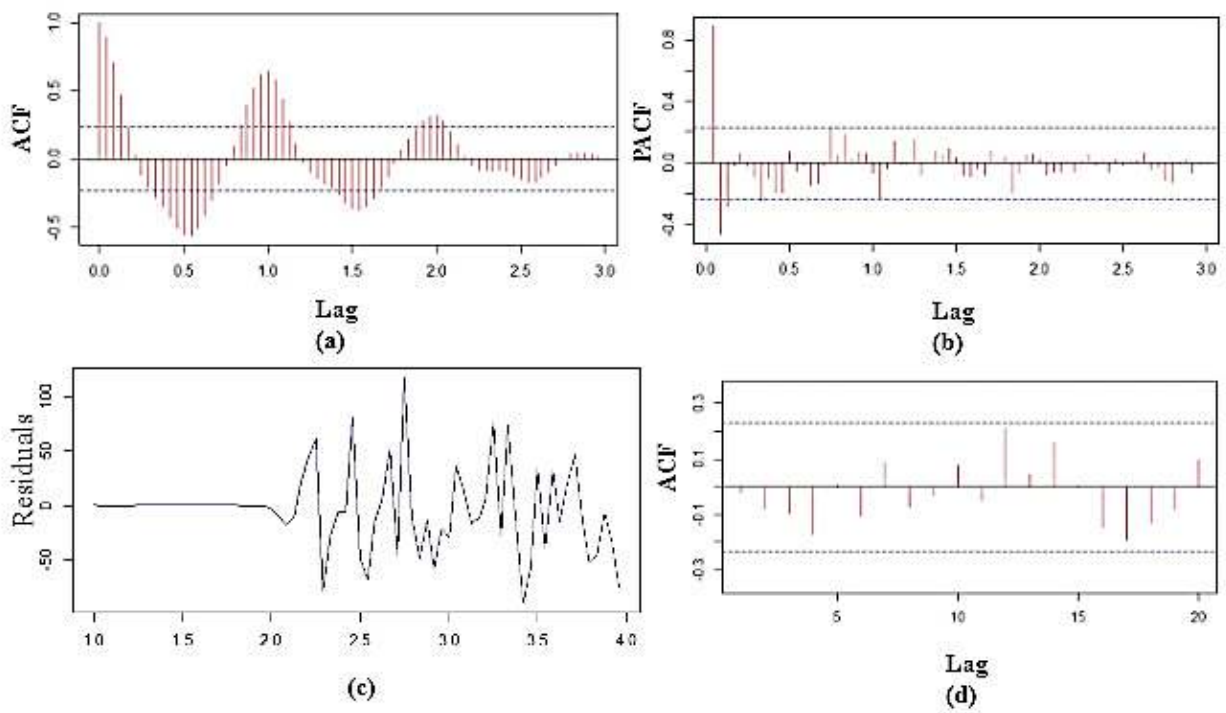

Fig. 6 (a) ACF for three-day traffic count data, (b) PACF for three-day traffic count data, (c) Residuals of the model, (d) ACF of the residuals from model fit

These inferences indicate that the time series day-wise data could be modelled using SARIMA. All the possible combinations of models were implicitly checked using R-software for minimum AIC and the best-prediction model was selected for the given time series. Based on this, the model ARIMA (2, 1, 0) (1, 1, 0) [24], was considered best for day-wise data. Once the possible models and their corresponding orders were found, the next step of model estimation is performed using R- software. For the selected model, the regression coefficients are significant, since the $|Z|$ is found to be more than 1.96, for all estimated coefficients. The next step in model-fitting is diagnostics. It starts with the analysis of the residuals as well as model evaluations. The residuals of the model are as shown in Fig.6c. It could be also possible to inspect the sample autocorrelations of the residuals, for any patterns or large values. ACF of residuals from the model fit is given in Fig. 6d, inspection of this plot shows no obvious patterns, and all the values were within the limits. Hence, the model appears to fit well. Similarly, SARIMA models were also applied for different input traffic count intervals such as 5-minute, 10-minute, and 15-minutes, based on this, traffic count for next day was predicted, with same time intervals.

4.2 Monte Carlo Simulation (MCS)

The dataset was divided into two different subsets, training set for model development, and testing set for validation. The 24 hours' vehicle count data was retrieved in every five-minute interval for six consecutive days, which was used for model development. Model was validated using the observed count of the next (seventh) day. Model development dataset was then analyzed for probability distribution functions viz. normal distribution and log normal distribution. Kolmogorov-Smirnov and Chi square tests were done to check goodness-of-fit of the fitted distributions. The statistics obtained for normal distribution (p-value: 0.76) and log normal distribution (p-value: 0.13), shows that both probability distributions fit well. Nevertheless, the goodness-of-fit tests showed a better fit for normal distribution. Using the mean and standard deviation of the fitted normal distribution, the new data set was produced with random probabilities. Similarly, for each hour, five-minute interval data for the next day was generated. A comparison of predicted values using MCS and observed traffic count frequency distribution is shown in Fig. 7.

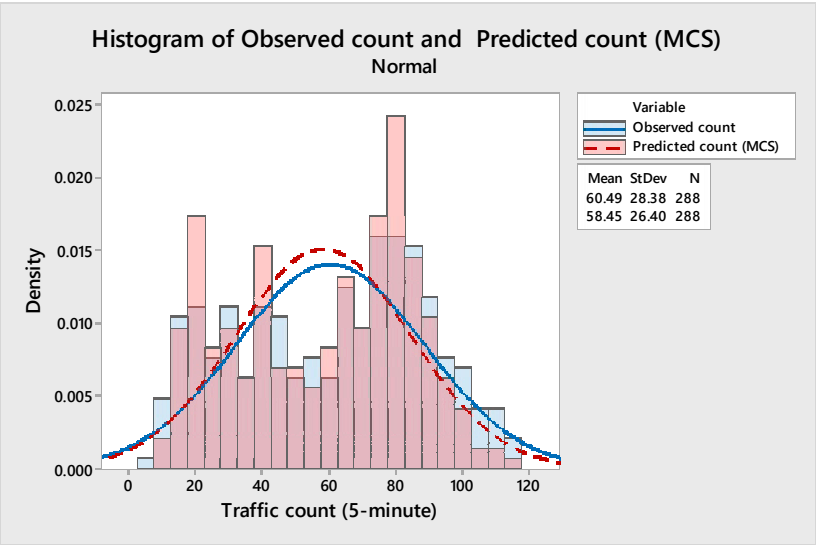

Fig. 7 Histogram of Observed count and predicted count for one-day, 5-minute interval, Location 1

4.3 Random Forest (RF) algorithm

Random forest ensemble learning technique is also used to obtain the short-term traffic count prediction, for this purpose regression model, developed based on random forest ensemble learning method is built using the 'randomForest' class in R software. The optimum number of decision trees ('ntree') and minimum number of observations per tree ('mtry') are fixed using 'tuneRF' function in R software. In this study, random forest was initially grown with 100 trees. To check whether the error decreases with a greater number of trees, plot between the number of trees and Out of the Bag error (OOB) is made, as depicted in Fig. 8. One of the sample decision trees is shown in Fig. 9. It was observed that the ensemble method with 100-trees was found have the minimum error as shown in Fig. 8.

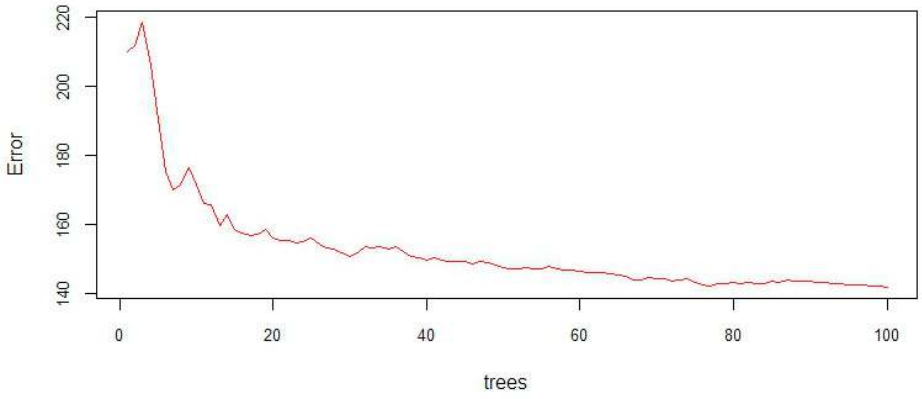

Fig. 8 Random Forest Error Vs No: of trees

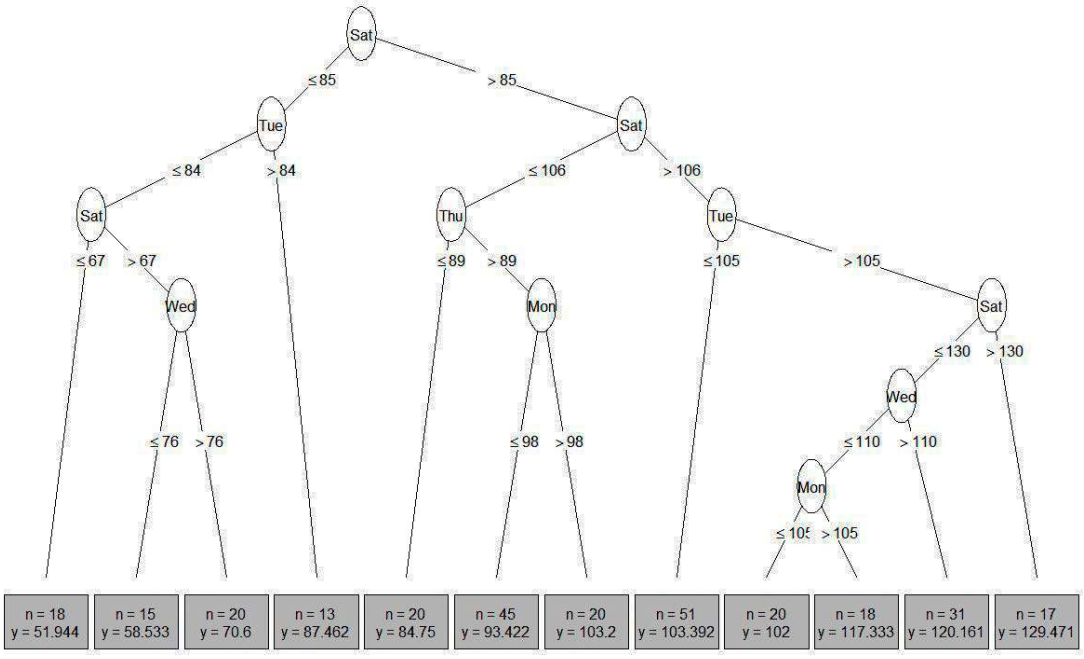

Fig. 9 Sample decision tree for 5-min interval traffic count prediction from R software

4.4 K-nearest neighbor (KNN) Technique

Present study also attempts to predict traffic count using KNN non-parametric regression. The total data collected was used as historical database for KNN algorithm. The algorithm was prepared using R software. For the search procedure, Euclidean distances (L2 in Equation (3)) were considered. Next step was the forecast generation, for which simple average of the dependent variable values of the nearest neighbours (Equation (4)) were used. Optimum K value can be found out using R-software for minimum error. Accuracy and K-value are plotted as shown in Fig. 10. From this graph, K value was decided as 6.

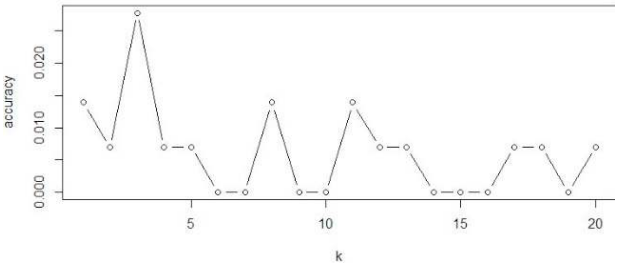

Fig. 10 Accuracy of prediction Vs K Nearest Neighbors considered

5. PERFORMANCE EVALUATION OF PREDICTION MODELS

In the performance evaluation stage, the performance of the proposed models was compared over varying time intervals of count. For this purpose, the collected data from 2 locations were used for training and validation. Performance of each prediction model was compared with a base method, namely Simple Averaging Technique (SAT), which is the average of the previous days' traffic count. The validation of the prediction models can be performed using Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) between the predicted and observed values. Comparison of MAPE and MAE was carried out in following sections for time series models SARIMA, MCS, RF and KNN algorithms. It can be inferred, in general, all the prediction models proposed are widely applied to Location 1 and 2 data for 5 minutes, 10 minutes, and 15 minutes intervals.

Fig. 11(a) and 11(b) show the MAPE and MAE for different methods for each time interval at Location 1 and Fig. 11(c) and 11(d) show the same results for Location 2. The MAPE of SAT, MCS, RF, SARIMA, and KNN are compared for both the locations for all three-time intervals. It is clear from the figures, for 5-minute interval prediction the MAPE ranges from 8-16%, for 10-minute interval prediction it is 7-12%, and for 15-minute interval prediction it is 5-11%, which is adequate in most of the ITS applications. According to Lewis' scale of interpretation of estimation accuracy (2), any forecast with a MAPE value of less than 10 % can be considered highly accurate, 11–20 % is good, 21–50 % is reasonable and 51 % or more is inaccurate. Hence it can be seen that the results are highly accurate.

When compared to SAT method, all the algorithms perform well for all the time intervals and for both the locations. In most of the cases KNN outperformed the rest of the models. Machine learning approaches like KNN and RF show better accuracy for prediction in all the time intervals at both locations. Fig. 11(b) and 11(d) show the MAE for different approaches for location 1 and 2, respectively. For location-1, all the approaches, the maximum MAE obtained for 5-minute interval is 8, where the average observed count from the field for same interval of the day was 60. Similarly, for location-2 is 11 and 90, respectively. Same for 10-minute interval, for location-1 is 14 & 120, and for location-2 is 16 & 196 respectively. In case of 15-minute interval the values for location-1 is 17 & 181 and for location-2 is 22 & 319, respectively. This demonstrates credibility of different approaches in predicting traffic toll-way counts.

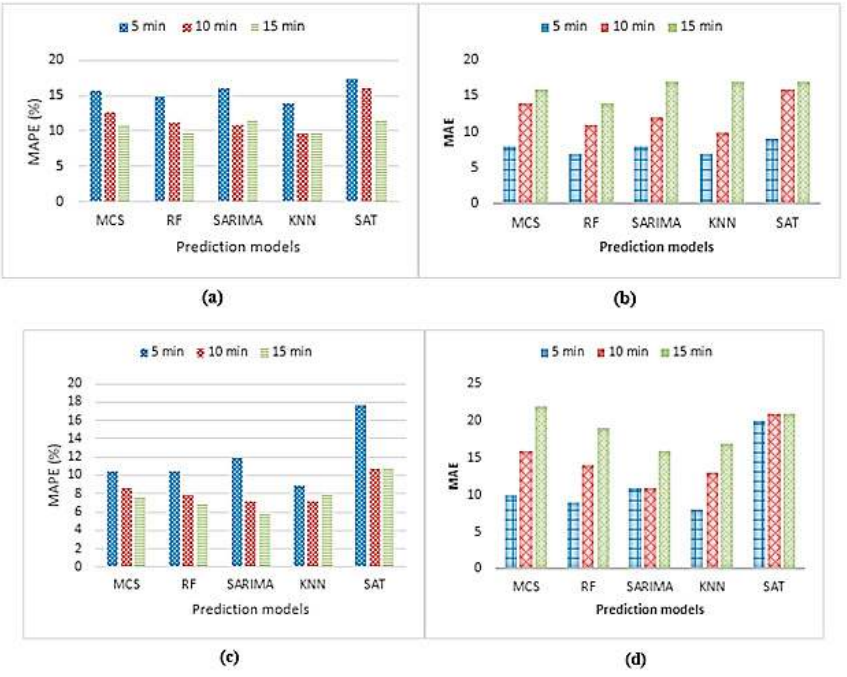

Fig. 11 (a) MAPE & (b) MAE for (Location 1), (c) MAPE & (d) MAE for (Location 2) (different approaches for all prediction intervals)

Fig. 12 shows the plot between observed traffic count and predicted traffic count for a day, using 5-minute interval data. Additionally, Fig. 13 and 14 show the comparison of observed and predicted counts for all the aforesaid models using 10-minut and 15-minute, respectively. From these figures it is clear that, all the observed and predicted counts at peak and off-peak hours are well matched and found to be highly competitive.

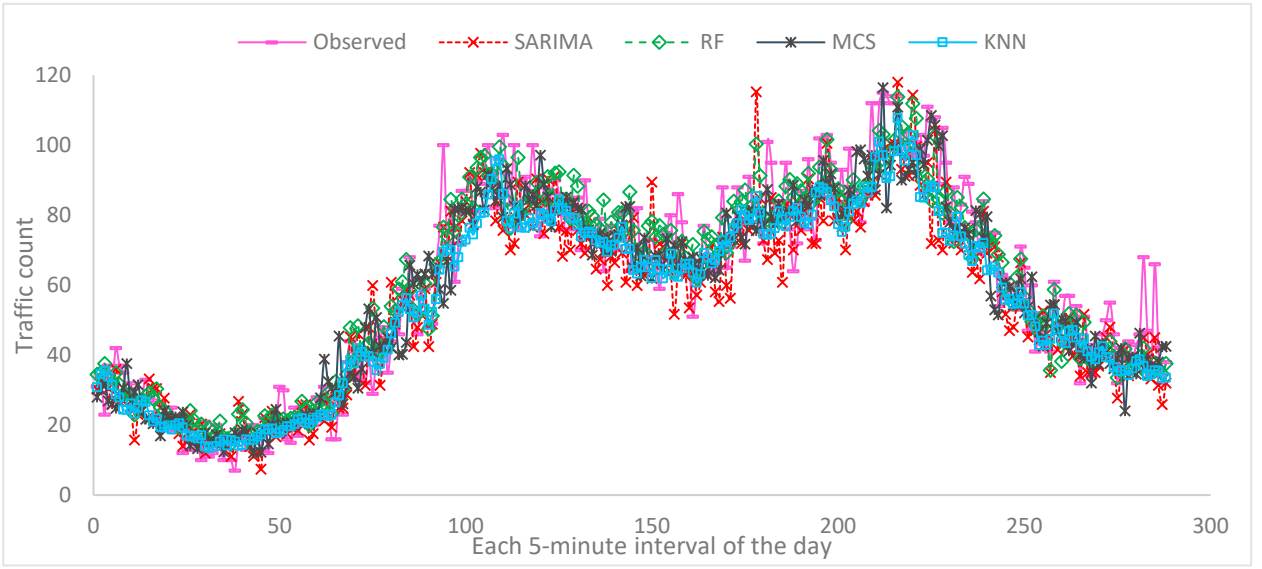

Fig. 12 Observed traffic count Vs Predicted traffic count for a day (5-minute interval) for SARIMA, RF, MCS and KNN respectively (Location 1)

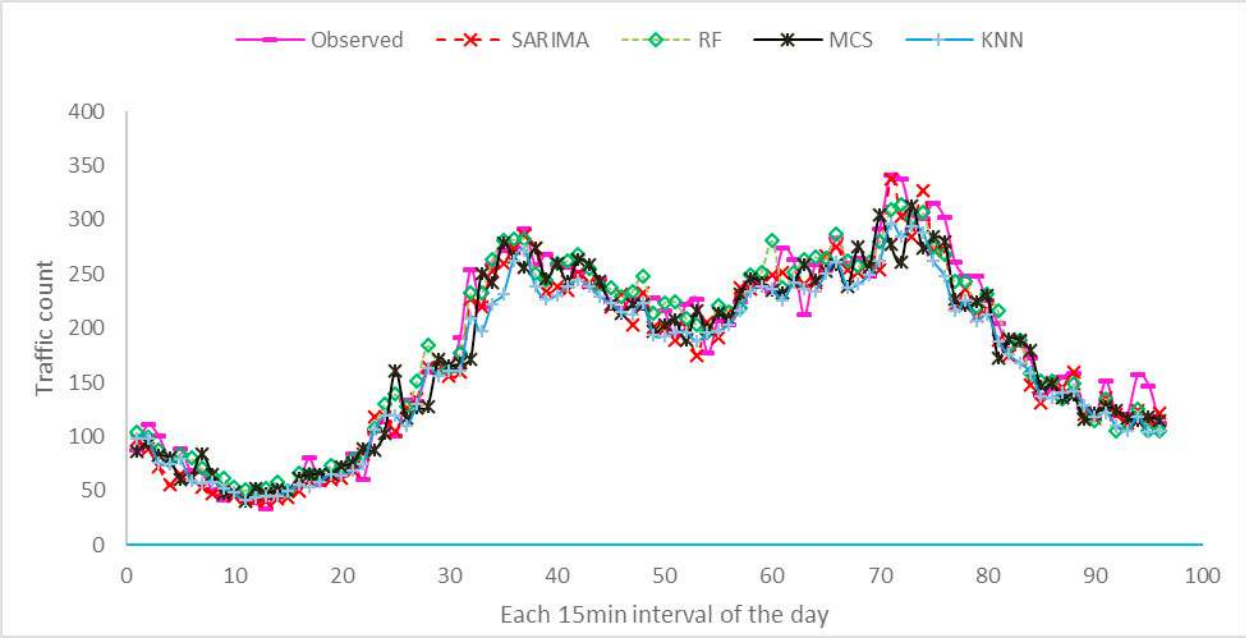

Fig. 13 Comparison of observed Vs Predicted traffic count for a day (10-minute interval) for SARIMA, RF, MCS and KNN (Location 1)

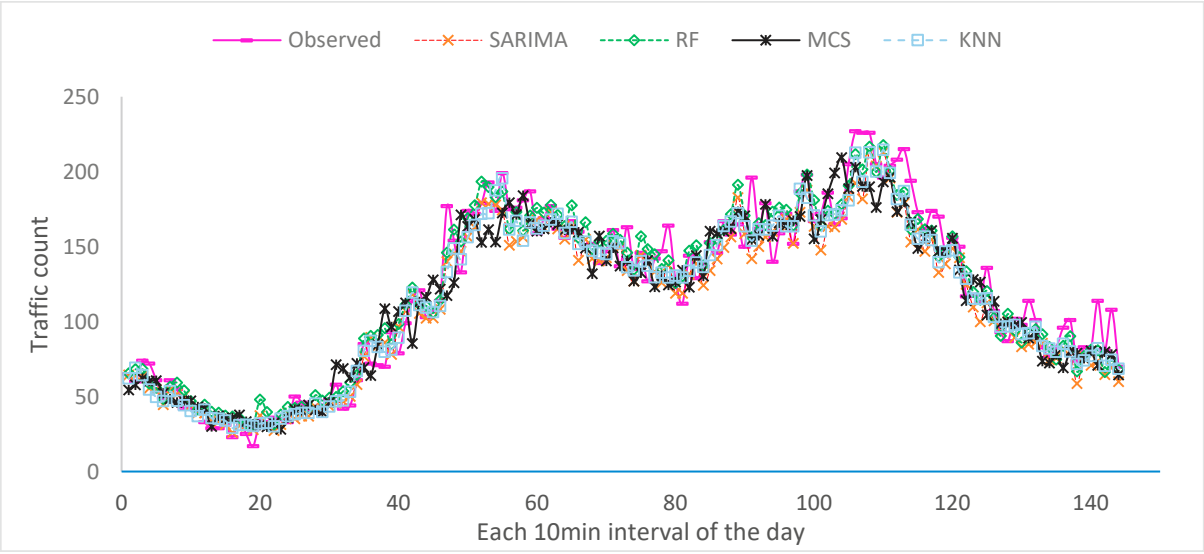

Fig. 14 Comparison of observed Vs Predicted traffic count for a day (15-minute interval) for SARIMA, RF, MCS, and KNN (Location 1)

5.1 Real-time traffic count prediction

To check the model accuracy and practical applicability of the prediction, short term traffic count prediction using real time data was also demonstrated. For this purpose, using the developed SARIMA model for location 2, the historic

data (previous 3-days traffic count data as input) was used along with real time data until the time of prediction on April 07, 2017. Both the morning (8-11am) and evening (5-8pm) peak periods of the day were predicted, as shown in Fig. 13(a) and 13(b), respectively. From Fig. 13(a) and 13(b), it may be noted that SARIMA time series technique is able to predict traffic counts at toll-plaza well, particularly for evening and morning peak hours with MAPE of less than 10%, can be considered highly accurate. This practical application of real-time toll-count performing accurate predictions, can be very inspiring for its actual implementation in field. This can be also really helpful for authorities to monitor even the vehicular growth at regional/national level, based on the historical time series data. This, eventually can be useful for studying even the traffic operations and dynamics on toll-roads.

Fig. 15 Comparison of observed and predicted count during (a) morning peak hours, (b) evening peak hours

5.2 Daily traffic count prediction for one month

By considering practical applicability of these prediction techniques in toll plazas, an effort has been made for the prediction of daily traffic count for one month. Inputting total daily traffic count of previous four months, next month daily traffic count has been predicted, which can be used to estimate monthly revenue in advance. Four consecutive months (April, May, June and July, 2015) daily traffic count data from second location was used for the model development and next one month (August 2015) was kept for validation. The prediction was performed using MCS and time series modelling. By considering the seasonality of the model ARIMA (2, 1, 0) (0, 1, 1) [30] was developed and is compared with ARIMA (5,1,1). Fig. 16 shows the observed and predicted count for the month of August, 2015 at location 2. MAPE of 9.56%, 10.91% and 13.01 % is obtained for MCS, ARIMA and SARIMA models respectively. The obtained accuracy is adequate for its practical applications.

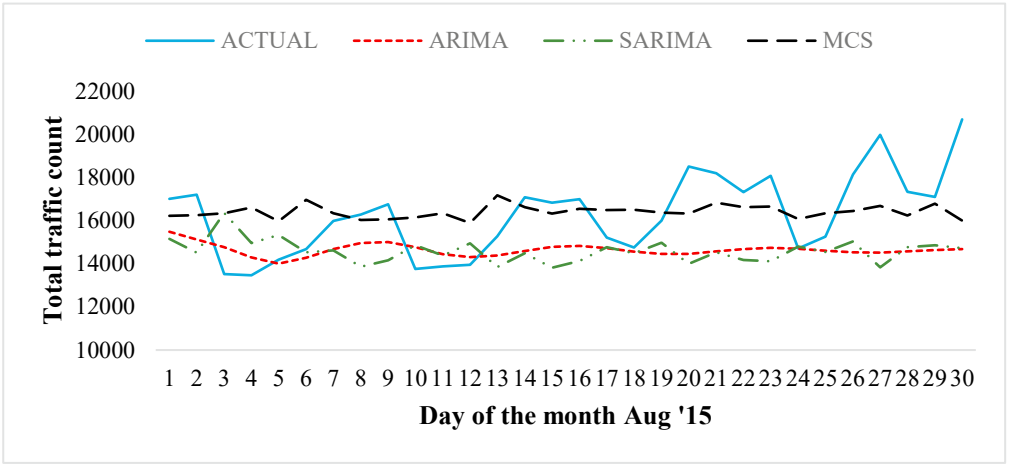

Fig. 16 Comparison of observed Vs Predicted traffic count for a month (Location 2)

6. CONCLUSIONS

The present study is focussed on toll way-count predictions using different techniques such as time series, parametric simulation based MCS algorithm, non-parametric machine learning algorithms such as KNN and RF. Keeping in view, this is going to be an imperative step in optimizing toll-way operations, as forecasted toll-way count is a mandatory input variable for any policy decisions for toll way operations, particularly for ATMS component in deployment of ITS at toll way. With this motivation, using the data collected from two different toll-plazas in India, four models were developed for short-term traffic count prediction.

To start with, temporal patterns of traffic counts were studied. A strong seasonal pattern with seasonality of 24 hours was observed in both the data sets. Based on this, SARIMA models were developed. In SARIMA models, previous 3 days' traffic count was used as input for predicting the next day (24 hour ahead) traffic count. SARIMA performed better for 10-minute input intervals of traffic count. MCS is also found to be highly competitive with SARIMA and it produces better results in most of the cases. But, it requires good amount of input data (normally distributed) as compared to SARIMA. Using MCS, there was not much variation in model performance for different input intervals. RF shows better performance than MCS and SARIMA. But in most of the cases, KNN outperformed the rest of the models. In the case of location-1 and location-2, one-week and one-month historic data were available for every 5-minute interval, respectively. Using KNN, less error was observed in location-2 prediction, than that of location-1, since KNN is a data driven approach. Nevertheless, as demonstrated, KNN can be one of the best approaches for predicting toll-way counts. Also, the present study explored the approaches to predict daily traffic count for one-month period. The results presented in this research work hint that many potential practice-oriented applications can be developed using demonstrated algorithms. Toll-way operations and other relevant logistics, such as number of channels, number of lanes, to be operated and deployment of man-power for specific time period/hour of the day and day of the week can be planned more precisely using precise estimate of toll-count in real-time.

References

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

BRIANL Smit. Michael J Demertsky, 2003. Traffic Flow Prediction: Neural Network Approach[J]. Washington, D.C: Transportation Research Record 1453,TRB.

Cheah, C.Y., Liu, J., 2006. Valuing governmental support in infrastructure projects as real options using Monte Carlo simulation. Construction Management and Economics, 24(5), pp.545-554.

Ghosh, B., Basu, B. and O'Mahony, M., 2009. Multivariate short-term traffic flow forecasting using time-series analysis. IEEE Transactions on Intelligent Transportation Systems, 10(2), pp.246-254.

Hamner, B., 2010. Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (pp. 1357-1359).

Jha, K., Sinha, N., Arkatkar, S.S. and Sarkar, A.K., 2016. A comparative study on application of time series analysis for traffic forecasting in India: prospects and limitations. CURRENT SCIENCE, 110(3), pp.373-385.

Kumar, S.V., Vanajakshi, L., 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. European Transport Research Review, 7(3), pp.1-9.

Leshem, G., Ritov, Y., 2007. Traffic flow prediction using Adaboost algorithm with random forests as a weak learner. Proceedings of World Academy of Science, Engineering and Technology (Vol. 19, pp. 193-198).

Lippi, M., Bertini, M., Frasconi, P., 2013. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. IEEE Transactions on Intelligent Transportation Systems, 14(2), pp.871-882.

Mishraa, R., Singhb, A.P., Sharmac, A., Sarkard, A.K., 2015. Short term traffic prediction using Monte Carlo simulation. 3rd Conference of Transport Research Group India.

Muthulakshmi, D., S., Janani, 2015. Design RFID tag and implementation of RSSI based automatic toll Connection. International journal of innovative research in electrical, electronics, instrumentation and control engineering Vol. 3.

R. Keith Oswald, Dr. William T. Scherer, Dr. Brian L. Smith., 2000. Traffic Flow Forecasting Using Approximate Nearest Neighbor Nonparametric Regression. A Research Project Report, UVA Center for Transportation Studies.

Raychaudhuri, S., 2008. Introduction to Monte Carlo Simulation. Simulation Conference, WSC 2008. Winter (pp. 91-100). IEEE.

Shumway, R.H., Stoffer, D.S., 2000. Time series analysis and its applications. STUDIES IN INFORMATICS AND CONTROL, 9(4), pp.375-376.

Smith, B.L., Williams, B.M., Oswald, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies, 10(4), pp.303-321.

Ta-Yin Hu, Wei-Ming Ho, 2010. Travel Time Prediction for Urban Networks: The Comparisons of Simulation-based and Time-Series Models, Department of Transportation and Communication Management Science.

Wang, S., Chen, W., 2011. Transportation Volume Forecast Methods Based on Time-Series Model. In ICTE 2011 (pp. 25-30). ASCE.
Wang, X., An, K., Tang, L., Chen, X., 2015. Short Term Prediction of Freeway Exiting Volume Based on SVM and KNN. International Journal of Transportation Science and Technology, 4(3), pp.337-352.