A comparative study on application of time series analysis for traffic forecasting in India: prospects and limitations

Kartikeya Jha[1], Nishita Sinha[2], Shriniwas S. Arkatkar[3,]* and Ashoke K. Sarkar[4]

[1]Zachry Department of Civil Engineering, Texas A&M University, College Station 77840, TX, USA
[2]Department of Agricultural Economics, Texas A&M University, College Station 77843, TX, USA
[3]Department of Civil Engineering, National Institute of Technology, Surat 395 007, India
[4]Department of Civil Engineering, Birla Institute of Technology and Science, Pilani 333 031, India

Modelling of growth trend and improvement in forecasting techniques for vehicular population has always been and will continue to be of paramount importance for any major infrastructure development initiatives in the transportation engineering sector. Although many traditional as well as some advanced methods are in vogue for this process of estimation, there has been a continuous quest for improving on the accuracy of different methods. Time-series (TS) analysis technique has been in use for short-term forecasting in the fields of finance and economics, and has been investigated here for its prospective use in traffic engineering. Towards this end, results obtained from two other traditional approaches, namely trend line analysis and econometric analysis, have also been collated, underlining the better results obtained from TS analysis. A regression model has been developed for predicting fatality rate and its results have been compared with those from TS analysis. Based on the incentive provided by reduced errors obtained from using increasing number of data points for model-building, forecasting has been done for the year 2021 using time-series modelling. With most of the datasets used and locations analysed for forecasting, the TS analysis technique has been found to be a useful tool for prediction, resulting in lower estimation errors for almost all the cases considered. It has also been inferred that the proximity of the forecasting window to the sample dataset has a noticeable effect on the accuracy of time-series forecasting, in addition to the amount of data used for analysis.

Keywords: Regression model, time-series analysis, traffic forecasting, transportation engineering.

Traffic forecasting

IN essence, traffic volume forecasting is the process of estimating the number of vehicles or people likely to use different transportation facilities in the future. For instance, a forecast may estimate the number of vehicles on a planned road or bridge, the expected ridership on a railway/metro line, the number of passengers visiting an airport, or the expected future traffic levels for the whole country. This process begins with the collection of data on current traffic. Depending upon the specific requirement of analysis, the traffic data are combined with other known data, such as population and economic growth rates, employment rate, trip rates, travel costs, etc. to develop a traffic demand model for the current situation. Combining this with predicted data for population, employment, etc. results in estimates of future traffic, typically estimated for each segment of the transportation infrastructure in question, e.g. for each roadway stretch or railway station that falls under the scope of facility.

Need for traffic forecasting

Knowledge of future traffic flow is an essential input in the planning, implementation and development of a transportation system. It also helps in its operation, management and control[1]. It is required to start the planning and/or development phase of any major transportation project initiatives. Being the first step in defining the scope and geometry of such projects, sometimes forecasting even helps us know whether a project is needed at all. Forecasting is necessary for doing relevant economic analysis[2]. It can also be used for other purposes such as corridor planning, systems planning, air quality analysis, safety analysis and other such special projects. Inaccuracies in traffic volume forecasts are responsible for the additional costs associated with over and under design[3]. The costs associated with an under-designed project arise when an additional project must satisfy the original inadequacies. Extra materials, labour and additional right-of-way attainment add to the cost of an over-designed project. Efficiency of traffic forecasting depends mainly on the size of average daily traffic. In general, the smaller the average daily traffic, the larger is the error in traffic forecasting. The major reasons for these errors can be:

- The changing traffic patterns in future, specifically induced demand effect[4,5] and rebound effect[6].

*For correspondence. (e-mail: sarkatkar@gmail.com)

- Traffic impacts due to development, majorly due to change in land-use patterns[7].
- Unforeseen and unaccounted socio-economic changes[8].
- Construction of new roads, diversions, etc.

Literature review

The literature review for this work comprises of the study of available literature on the methods previously used for traffic forecasting, their challenges, scope for improvement and then the study of more recent, contemporary approaches to forecasting, especially with reference to time-series (TS) analysis. In the Indian context, the past research work has mainly concentrated on trend line analysis[9,10]. Here the traffic volume levels for the country have been predicted using a linear relationship between the GNP of a country and the total vehicular population. On the same lines, a project feasibility report[11] on six-laning of NH-2 from Delhi to Agra prepared by CES for NHDP, India, elaborates a combination of trip generation models and trend line analysis using NSDP instead of GNP for different corridors lying in the scope of this project (see Appendix 1 for abbreviations used). Study of more contemporary areas of research focuses mainly on the TS analysis. While Bhar and Sharma[12] deal with the applications and nuances of TS analysis (exemplified with the use of the SPSS software, Nihan and Holmesland[13] stress on the basics of TS modelling. Approximate nearest neighbor nonparametric regression method has been discussed by Oswald *et al.*[14].

Time-series models have been extensively used in traffic forecasting for their simplicity and strong potential for on-line implementation (see, for example, refs 13, 15–26). ARIMA is a variant of regression modelling used to work with time-series data in statistics and econometrics. This technique and its elements are discussed in detail in the 'methodology' section. Levin and Tsao[16] used an ARIMA (0, 1, 1) model, which is essentially the same as exponential smoothing model. Their comparison with an ARIMA (0, 1, 0), or the random walk model, showed that the ARIMA (0, 1, 1) model performed better. Ahmed and Cook[17] proposed an ARIMA (0, 1, 3) model. They compared the results with those obtained by other models like double exponential smoothing, simple moving average and exponential smoothing with adaptive response. The authors found that ARIMA (0, 1, 3) gave better forecasts than the other three models. Nihan and Holmesland[13], and Moorthy and Ratcliffe[19] also applied the Box–Jenkins method to produce the short-term forecasts of daily flows. Hamed *et al.*[21] investigated the use of time-series models for predicting arterial traffic flow. The data consisted of 1 min traffic flow for morning peak hours (6 : 30–8 : 15 a.m.). They fitted several ARIMA models after using first-order ordinary differencing. Results showed that the ARIMA (0, 1, 1) model best described the data. Tang and Lam[27] applied the Box–Jenkins ARIMA model using historical and current-year partial daily flow developed for short-term prediction of daily flows using Hong Kong data. A study by Smith *et al.*[28] compared parametric and nonparametric models for traffic flow forecasting. The study showed that prediction under seasonal ARIMA, a parametric modelling approach to time series, outperforms other nonparametric approaches like regression based on heuristically improved forecast generation. However, the results did indicate that in cases when the implementation requirements of seasonal ARIMA models cannot be met, using nonparametric regression coupled with heuristic forecast generation methods is preferred.

Tang *et al.*[29] adapted time-series, neural network, nonparametric regression, and Gaussian maximum likelihood methods to develop models for predicting traffic volumes by day of the week, month and AADT for the entire year. Analysis was conducted based on historical traffic data Hong Kong from 1994 to 1998. The daily flows estimated by the four models were used to calculate the AADT for the year of 1999. The results from the four models were compared and the Gaussian maximum likelihood model appeared to be the most promising and robust among them for extensive applications to provide short-term traffic forecasting database for the whole territory of Hong Kong. Chandra and Al-Deek[30] found that the past values of upstream as well as downstream stations influence the future values at a station and therefore can be used for prediction. Further, it was also found that a vector autoregressive model is appropriate and better than the traditional ARIMA model for prediction at these stations. Although a number of methods can be adopted for traffic volume forecasting depending on the specific situation at hand, for the present analysis one of the more recent approaches, i.e. TS analysis was chosen for comparison with other traditional methods.

Objective and scope

This article attempts to highlights the usefulness of TS analysis in traffic forecasting by underlining the lower values of estimation errors found with this method when compared to two other methods – trend line analysis and econometric regression analysis cited from a previous work[31]. The results of analyses with increasing data points have been compared for relative levels of accuracy. Moreover, a regression equation has been arrived at for predicting fatality rate (number of persons killed in road accidents) in India and results from this equation have been compared with those from TS analysis to check for relative accuracies.

This exercise is only the first step in developing an insight into the choice of the best-suited method to estimate future traffic levels in a country which, as has been

Table 1. Data for total vehicular population and other variables (1971–2006)

Year	Vehicular population (in ‘000)	Persons killed (D)	Population (million) (P)	Year	Vehicular population (in ‘000)	Persons killed (D)	Population (millions) (P)
1971	1865	15,000	560.2675	1989	16,920	50,700	832.535
1972	2045	16,100	573.1299	1990	19,152	54,100	849.515
1973	2109	17,600	586.2198	1991	21,374	56,400	866.53
1974	2327	17,300	599.6427	1992	23,507	57,200	882.821
1975	2472	16,900	613.459	1993	25,299	60,700	899.329
1976	2700	17,800	627.6324	1994	26,464	64,000	915.697
1977	3260	20,100	642.1336	1995	30,125	70,600	932.18
1978	3614	21,800	656.9406	1996	33,786	74,600	948.7589
1979	4059	22,600	672.0209	1997	37,332	77,000	965.4282
1980	4521	24,600	687.332	1998	41,368	79,900	982.1825
1981	5391	28,400	702.8212	1999	44,875	82,000	999.016
1982	6055	30,700	718.4256	2000	48,857	78,900	1015.923
1983	6973	32,800	734.072	2001	54,991	80,900	1032.473
1984	7949	35,100	749.6769	2002	58,924	84,600	1048.641
1985	9170	39,200	765.147	2003	67,007	85,900	1064.399
1986	10,577	40,000	781.893	2004	72,718	92,600	1079.721
1987	12,618	44,400	798.68	2005	81,501	94,900	1094.583
1988	14,818	46,600	815.59	2006	89,618	105,700	1109.811

Source: Ministry of Road Transport & Highways, Government of India (www.morth.nic.in).

discussed, is imperative from many aspects. Due to data availability constraints the present analysis has been done for total vehicular population in India to enable the choice of appropriate methods for estimation at specific project level also. The primary data used have been cited from the MoRTH, Government of India (www.morth.nic.in). This has been reproduced in Table 1 for ready reference. To gauge the extent of data requirement of the TS method, the analysis was carried out with 22, 26, 30, and 35 years vehicular population data and respective errors in estimation were calculated.

Encouraged by the improvement in results obtained as more and more data were used for model creation and validation, and as we moved closer to the forecasting window, which is reflected in the constantly diminishing values of RMSE calculated by comparing predicted values with actual figures for the entire dataset, forecasting has been done for the year 2021 for total vehicular population and fatality rate in India. As suggested by Box and Jenkins[32], ideally at least 50 observations are required for performing TS analysis. Taking this into account, TS analysis was done on AADT data sourced from Performance Measurement System (PeMS)[33], California, USA, for a location in district 7 on Interstate-10(W) (Table 2). This analysis further establishes the potential of TS analysis as a promising alternative to traditional methods of forecasting. Overall, the article attempts to gauge the suitability of TS forecasting technique for traffic volume prediction. Given rich and varied data availability, the analysis can be extended to produce better understanding of this method and its application to project-level studies as well. Further, multivariate TS modelling can be explored for even better results if data availability meets the high requirements of TS analysis.

Methodology of analysis

Methods adopted

This work deals mainly with the TS analysis for forecasting. At the same time, for a similar dataset, the results obtained after analysis by this method have been compared with those obtained from two other methods – trend line analysis, where future traffic volume is predicted based on a linear relationship between traffic population and GNP; and econometric regression analysis, where traffic demand is seen as being dependent on chosen economic/demographic variables, as proposed by Jha *et al.*[31]. This method has also been used to compare with results for number of fatalities in road accidents from a regression analysis between fatality rate and vehicle ownership rate. A brief description of the TS method is given below:

Time series analysis: Time series is a set of observations ordered in time. This analysis deals with observations that are collected over equally spaced, discrete time intervals. As in this case, when observations are made for only one variable over time, it is called a univariate time series. The fundamental assumption for any TS analysis is that some aspects of past pattern will continue to affect the future values. Values of variables occurring prior to the current observation are called lag values. The primary difference between TS models and other types of models is that lag values of the target variable are used as predictor variables, whereas traditional models use other variables as predictors, and the concept of a lag value does not apply because the observations do not represent a chronological sequence. A time series is deterministic if its future behaviour can be exactly predicted from its past

Table 2. Monthly AADT data for location 'Lark Ellen' on I-10(W), California, USA

Month	Mainline (ML) AADT	Month	Mainline (ML) AADT	Month	Mainline (ML) AADT
July 2000	117,018	February 2004	120,744	September 2007	116,075
August 2000	117,170	March 2004	120,914	October 2007	116,076
September 2000	117,113	April 2004	120,896	November 2007	115,953
October 2000	117,099	May 2004	120,710	December 2007	115,650
November 2000	117,339	June 2004	120,672	January 2008	115,364
December 2000	117,462	July 2004	120,258	February 2008	115,341
January 2001	117,725	August 2004	119,920	March 2008	115,093
February 2001	118,230	September 2004	119,160	April 2008	115,009
March 2001	118,441	October 2004	118,394	May 2008	115,176
April 2001	118,147	November 2004	118,376	June 2008	115,351
May 2001	118,070	December 2004	121,533	July 2008	115,630
June 2001	118,020	January 2005	124,428	August 2008	115,670
July 2001	117,952	February 2005	125,687	September 2008	115,698
August 2001	118,524	March 2005	125,992	October 2008	115,596
September 2001	119,371	April 2005	126,534	November 2008	115,441
October 2001	119,828	May 2005	126,209	December 2008	115,472
November 2001	119,624	June 2005	126,098	January 2009	115,743
December 2001	119,469	July 2005	125,727	February 2009	116,064
January 2002	119,249	August 2005	125,454	March 2009	116,292
February 2002	118,846	September 2005	125,454	April 2009	116,346
March 2002	118,790	October 2005	125,431	May 2009	116,031
April 2002	118,425	November 2005	125,496	June 2009	115,870
May 2002	118,615	December 2005	125,283	July 2009	115,563
June 2002	118,910	January 2006	125,427	August 2009	115,462
July 2002	119,169	February 2006	125,349	September 2009	115,122
August 2002	119,426	March 2006	125,197	October 2009	115,156
September 2002	119,412	April 2006	124,803	November 2009	115,486
October 2002	119,364	May 2006	124,623	December 2009	115,616
November 2002	119,450	June 2006	123,341	January 2010	114,961
December 2002	119,397	July 2006	122,467	February 2010	114,377
January 2003	119,689	August 2006	121,037	March 2010	114,107
February 2003	119,904	September 2006	120,781	April 2010	113,668
March 2003	120,019	October 2006	120,445	May 2010	113,356
April 2003	120,451	November 2006	119,571	June 2010	112,959
May 2003	120,665	December 2006	118,978	July 2010	112,688
June 2003	120,594	January 2007	118,558	August 2010	112,579
July 2003	120,750	February 2007	117,770	September 2010	112,617
August 2003	120,717	March 2007	117,690	October 2010	112,282
September 2003	120,831	April 2007	117,817	November 2010	111,810
October 2003	121,107	May 2007	117,860	December 2010	111,340
November 2003	120,832	June 2007	117,484	January 2011	111,574
December 2003	120,931	July 2007	116,822	February 2011	111,668
January 2004	120,731	August 2007	116,424	March 2011	111,834

Source: http://www.pems.dot.ca.gov (PeMS, Caltrans; accessed on 2 April 2012).

behaviour. Otherwise the time series is statistical. The future behaviour of a statistical time series can be predicted only in probabilistic terms.

TS techniques can be used to develop highly accurate and inexpensive short-term forecasts. The Box and Jenkins methodology has been adopted and analysis has been done using the ARIMA approach[34]. The main rationale behind using the Box and Jenkins technique is that it has been shown to produce relatively accurate forecasts. The results from comparative studies conducted by Naylor *et al.*[35], and Nelson[36] showed that the Box and Jenkins model, albeit simpler, was more effective than other such contemporary econometric models. The basic limitation of this approach is its high data requirement. In case of traffic forecasting, it demands rich, reasonably accurate data spanning over a long time-frame so that there may be sufficient number of data points to model the situation appropriately. Because this is not always possible in the Indian context, there is scope for improvement if this approach is to be used to good effect in the future.

Analysis

As has been remarked before, there are two sets of TS analysis that have been performed here. Both are discussed sequentially hereafter.

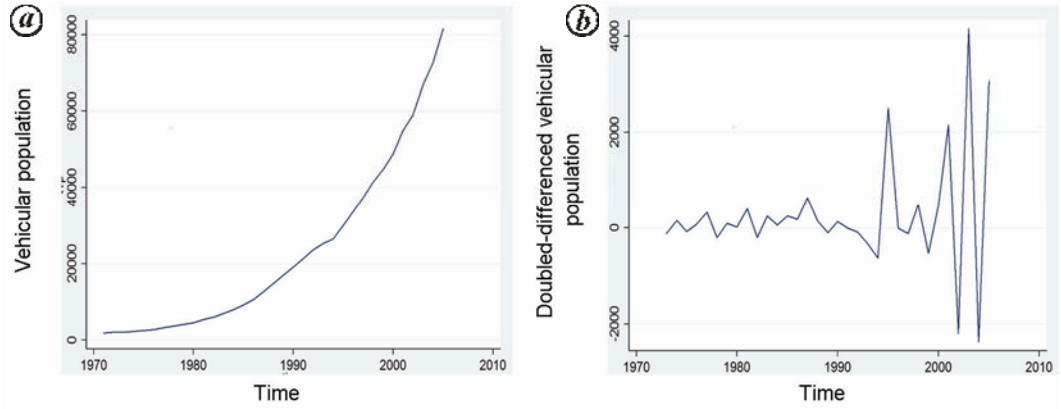

Figure 1. (***a***) Non-stationary data and (***b***) stationary data.

Analysis with MoRTH data: For the first case, the data used have been cited from MoRTH (Table 1). For univariate TS analysis, data from years 1971 to 2005 (35 years) have been used. The estimation has been done for the target year 2006. The Box and Jenkins methodology has been used and ARIMA technique has been adopted for analysis. The modelling has been performed on STATA. The following brief definitions will enable a better understanding of the TS analysis, and reasons behind selection of particular models for the same.

Box and Jenkins methodology: The original Box–Jenkins modelling procedure involved an iterative three-stage process of model selection, parameter estimation and model checking. The five broad steps include the following:

Checking for stationarity and transforming the dataset such that assumption of stationarity is reasonable: A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time or space. Consequently, parameters such as the mean and variance, if they exist, also do not change over time or position (Figure 1). Dickey Fuller and Philip Perron tests are performed to confirm stationarity of data used. A stationary series is relatively easy to predict because it is predicted that its statistical properties will be the same in the future as they have been in the past. The predictions for the stationary series can then be untransformed by reversing whatever mathematical transformations were previously used, to obtain predictions for the original series. Another reason for trying to stationarize a time series is to be able to obtain meaningful sample statistics such as means, variances and correlations with other variables. Such statistics is useful as descriptors of future behaviour only if the series is stationary. For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. And if the mean and variance of a series are not well-defined, so are its correlations with other variables.

Identification of the parameters of the model: ARIMA. Lags of the differenced series appearing in the forecasting equation are called 'auto-regressive' terms, lags of the forecast errors are called 'moving average' terms, and a time series which needs to be differenced to be made stationary is said to be an 'integrated' version of a stationary series. An ARMA model predicts the value of the target variable as a linear function of lag values (this is the auto-regressive (AR) part) plus an effect from recent random shock values (this is the moving average (MA) part). To get the order of AR and MA process, ACF and PACF are studied.

An autoregressive process is a function of lagged dependent variables and a moving average process a function of lagged error terms. An autocorrelation is the correlation between the target variable and lag values for the same variable. Correlation values range from –1 to +1. A value of +1 indicates that the two variables move together perfectly; a value of –1 indicates that they move in opposite directions. When building a TS model, it is important to include lag values that have large, positive autocorrelation values. Sometimes it is also useful to include those that have large negative autocorrelations. The partial autocorrelation is the autocorrelation of TS observations separated by a lag of time units with the effects of the intervening observations eliminated. The grey regions in the ACF and PACF plots in Figure 2 show points within two standard deviations (an approximate 95% confidence interval) from zero. If the autocorrelation/partial autocorrelation bar is longer than the marker (that is, it covers it), then the correlation should be considered significant.

Estimation of the parameters: There are two different ways in which a model can be estimated-maximum

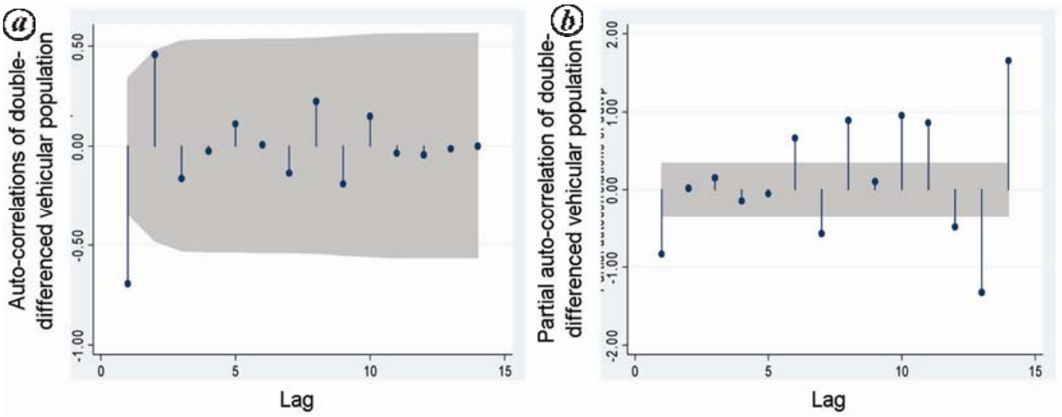

Figure 2. (***a***) ACF plot and (***b***) PACF plot.

likelihood estimation, and conditional maximum likelihood estimation. The first one uses numerical optimization techniques for estimation purpose and the latter is OLS regression. This analysis follows full maximum likelihood estimation. Based on minimum AIC and BIC which determine the parsimony of the model, the best models are selected. A parsimonious model is desirable because including irrelevant lags in the model increases the coefficient standard errors and therefore reduces the *t*-statistics.

Performing diagnostic checks: If the model is correctly specified, the residuals of the model should be uncorrelated. In other words, there should be a white noise. One way to test this is to get a Portmanteau test statistic. This is also called the white noise test. It indicates the absence of serial correlation or predictability. If the computed Q exceeds the value from the χ^2 table for some specified significance level, the null hypothesis that the series of autocorrelations represents a random series is rejected at that level. The *P*-value gives the probability of exceeding the computed Q, given a random series of residuals. Thus random residuals give small Q and high *P*-value. Results are considered better when the value of this probability is closer to 1.

Hence, following the above methodology, all models tested here have been qualified on the basis of the criteria mentioned below (in the same order): (i) The number of significant spikes as given by the ACF and PACF plots, (ii) High probability for white noise (non-auto-correlation of residuals), (iii) Low RMSE to ensure optimum accuracy and (iv) Among the models that gave favourable results for the above, further narrowing down was done for selection based on parsimony indicated by low AIC and BIC.

To assess the extent of data requirement for TS analysis, traffic population data for increasing number of years were first tried out. In all these cases, estimation has been done for the target year 2006. The respective observations for all these have been compiled as follows:

Table 3. TS test statistics for various prospective models with 30 years data

Model	*P* (white noise)	AIC	BIC	RMSE
ARIMA (0, 2, 0)	0.858	435.31	437.97	4086.56
ARIMA (6, 2, 0)	0.994	438.79	449.45	4183.30
ARIMA (7, 2, 0)	0.999	438.10	449.99	4549.73

With 15 years data (1971–85): No significant spikes were observed from the ACF and PACF plots. All the points were located well inside the grey region for both the plots. This indicated that the available data was insufficient to carry out TS analysis.

With 22 years data (1971–92): Although no significant spikes were observed for the ACF plot, the PACF plot showed some spikes going outside the grey region. ARIMA (0, 2, 0) and (7, 2, 0) models could be investigated during analysis. Both of these resulted in low probability for white noise and high RMSE, thus being insufficiently conclusive.

With 26 years data (1971–96): Both the ACF and PACF plots showed data consistency in terms of significance of spikes. ARIMA (2, 2, 0), (3, 2, 0) and (6, 2, 0) models were investigated during analysis. ARIMA (2, 2, 0) was selected for forecasting because it results in low values of AIC, BIC and RMSE and passes the white noise test favourably.

With 30 years data (1971–2000): Table 3 shows the various prospective models selected for analysis in this case along with relevant parameters. We observe that the results for the white noise test improve significantly and RMSE values reduce with respect to the previous sets of data used for 15, 22 and 26 years.

As discussed, the RMSE values in Table 3 have been calculated based on the error (deviation) between the actual and predicted total vehicular population in the period

Table 4. Component statistics for ARIMA (6, 2, 0) using 30 years data

ARIMA regression					Number of obs. = 28	
Sample: 1973–2000					Wald chi^2 (6) = 8.32	
Log likelihood = –211.3953					Prob > chi^2 = 0.2153	
D_2.TVP*		OPG	*z*	*P* > |*z*|	(95% conf. interval)	
TVP	Coef.	Std. Err.				
_cons	137.0784	63.29741	2.17	0.030	13.0178	261.1391
ARMA, ar						
L1.	–0.4126212	0.1997691	–2.07	0.039	–0.8041616	–0.0210809
L2.	–0.3930978	0.3802716	–1.03	0.301	–1.138416	0.3522208
L3.	–0.2361988	0.3605105	–0.66	0.512	–0.9427865	0.4703889
L4.	–0.4007741	0.8033748	–0.50	0.618	–1.97536	1.173812
L5.	–0.175019	0.4453681	–0.39	0.694	–1.047924	0.6978864
L6.	–0.4357381	0.507881	–0.86	0.391	–1.431167	0.5596905
σ	445.4789	49.55745	8.99	0.000	348.348	542.6097

*D_i. *X* refers to the *i*th consecutive difference for series of *X*; hence in the table, second difference of total vehicular population (TVP). Cons, Constant' Coef., Coefficient; Std. err., Standard error; Conf. interval, Confidence interval.

2001–06, which is the forecasting window. This applies to Tables 5 and 7 of the same nature also. Though the values of RMSE appear high in magnitude, the average RMSE is about 0.035% of the average base values for AADT. Also, the RMSE values increase as we move towards the extreme ends (left and right) of the dataset. In the middle portion, they remain at relatively lower levels. On the basis of the results obtained, ARIMA (6, 2, 0) was considered best suited for estimation based on high white noise, low AIC, BIC and RMSE values.

Table 4 shows significant test statistics for the ARIMA (6, 2, 0) model. A few key observations have been discussed to have a better interpretation of the descriptive statistics. The estimated coefficients (column 2 in Table 4) should be significantly different (distant) from zero. Significance of the AR and MA coefficients can be evaluated by comparing estimated parameters with the standard errors (column 3). The value (magnitude-wise) of the standard error should be less than that of the coefficient itself. In this case, except for lag 1 of AR component, the parameters could not be found to be significant at 5% level of significance, as the 95% confidence interval (last two columns in Table 4) included the point 0.

As the confidence interval was lowered, some other lags were found to be significant at 75% confidence interval. This could have been caused due to the less than adequate data, the extent of veracity of data source and method of data collection adopted for the data used for TS analysis in this case. Figure 3 shows the result of forecasting performed for the year 2006 using 30 years data. The dashed line in the plot shows the predicted values, while the solid line marks the actual values for total vehicular population for each corresponding year. The gap between these two, which is more apparent in the final time range (2000–06), indicates underestimation as the predicted values are lower than the actual ones.

Table 5. TS test statistics for various prospective models with 35 years data

Model	*P* (white noise)	AIC	BIC	RMSE
ARIMA (0, 2, 0)	0.003	565.97	568.96	1436.01
ARIMA (1, 2, 0)	0.958	540.33	544.82	866.39
ARIMA (1, 2, 1)	0.958	542.33	548.32	865.86

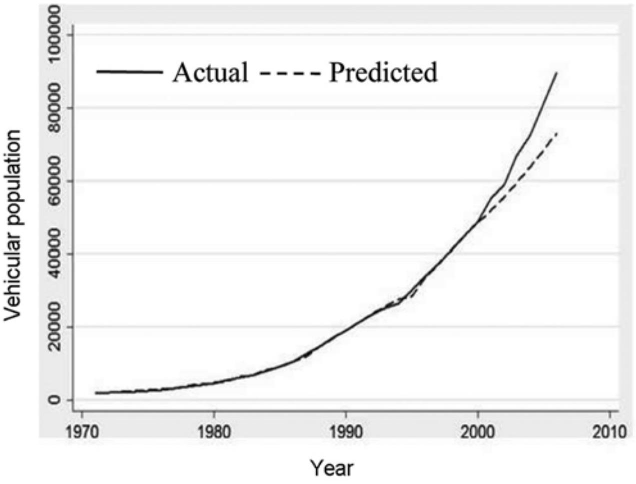

Figure 3. Actual and predicted values for ARIMA (6, 2, 0) using data for 30 years.

With 35 years data (1971–2005): The available data had to be differenced twice to achieve stationarity (a pre-requisite for TS analysis), as shown in Figure 1. The Dickey Fuller and Philip Perron tests were conducted to confirm stationarity. Table 5 provides the statistical results obtained for various models using this set of data. We observe that the RMSE values fall noticeably compared to the previous cases. This may have been caused due to two significant factors: (i) increase in the amount

Table 6. ARIMA (1, 2, 0) regression parameters

ARIMA regression					Number of obs = 33	
Sample: 1973–2005					Wald chi^2 (1) = 85.31	
Log likelihood = –267.1671					Prob > chi^2 = 0.0000	
D_2.TVP		OPG	z	$P > \|z\|$	(95% conf. interval)	
TVP	Coef.	Std. Err.				
_cons	226.9305	113.9699	1.99	0.046	3.553639	450.3074
ARMA						
ar, L1.	–0.8030058	0.0869377	–9.24	0.000	–0.9734006	–0.6326109
σ	781.6101	114.2785	6.84	0.000	557.6283	1005.592

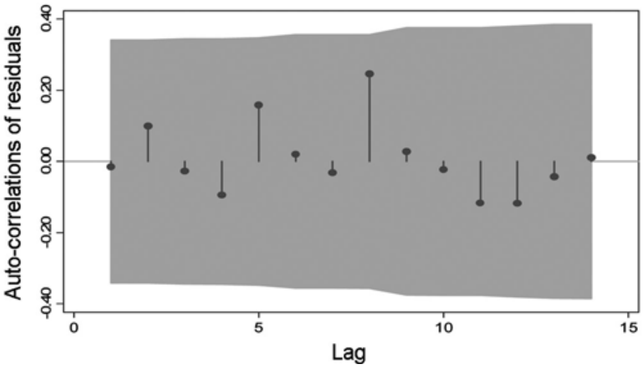

Figure 4. ACF plot for residuals of the model ARIMA (1, 2, 0).

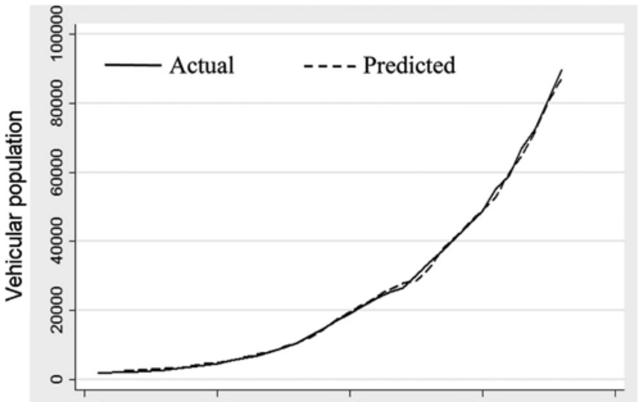

Figure 5. Actual and predicted values for ARIMA (1, 2, 0) using data for 35 years.

of data used for modelling, and (ii) Increase in the proximity of the forecasting window (year 2006) to the sample data used (1971–05) compared to cases taken up earlier.

Table 6 shows the ARIMA regression parameters for ARIMA (1, 2, 0) model. Portmanteau test for white noise gave Portmanteau (Q) statistic as 6.326, which is less than the critical value of 23.7 at 5% level of significance. Also, probability $>\chi^2$ (14) is 0.9583, which is very close to 1. The plot of ACF for the residuals of this model also suggests that they are non-auto-correlated, as none of the spikes goes beyond the 2σ region (Figure 4). Considering these results and their relative significance, for this case ARIMA (1, 2, 0) was considered as the best suited for modelling. Figure 5 shows the plot for actual and predicted values obtained for ARIMA (1, 2, 0) using data for 35 years (1971–2005). Keen observation shows that the gap which was evident in case of 30 years data has closed in as the forecasting window moves closer to the sample data. This may have been caused due to the fact that data in the window 2000–05 were used for modelling itself and the forecasting window consisted of the year 2006 only.

In order to ascertain whether it is the quantity (volume) of data used for the analysis or the proximity of dataset to the forecasting window that reflects on the accuracy of results from the analysis, the TS analysis was done on the same dataset, this time using data till 2005, but for 25 and 30 years respectively, for forecasting figures for the year 2006. Thus for both these datasets, the forecasting window is equally close to the dataset, while the amount of data (number of data points) used differs. The following is the description of the analyses done:

With 25 years data (1981–2005): ARIMA(1,2,0) was chosen out of four prospective models for forecasting based on high white noise probability, low AIC, BIC and RMSE values.

With 30 years data (1976–2005): ARIMA (1, 2, 0) was chosen out of four prospective models for forecasting based on high white noise probability, low AIC, BIC and RMSE values.

Also, forecasting was done for the year 2021 using data from 1971 to 2006. ARIMA (0, 2, 0), (0, 2, 1), (1, 2, 0), (1, 2, 1) and (1, 2, 2) models were investigated for modelling and based on the results for relevant statistics for these models, ARIMA (1, 2, 1) was considered the best. The value for total vehicular population predicted by this model for the year 2021 is 236,269,000.

To further exhibit the potential of TS modelling, a comparative analysis has been done between results from two methods – regression and TS analysis. Both these analyses use the same dataset to enable suggestive comparison of the results. As suggested by Vishwas *et al.*[37], a regression model has been built between number of deaths (D) per 10,000 vehicles (N) and vehicle ownership (number of vehicles in 10,000 per million population), that is D/N versus N/P using data from 1971 to 2000 for predicting future fatality rate. Table 1 shows the relevant data. While Figure 6 shows the best-fit curve. The equation that emerges from such a regression analysis is:

$$\frac{D}{N} = 43.858\left(\frac{N}{P}\right)^{-0.564} \quad (R^2 = 0.993).$$

Using the same dataset (1971–2000), TS models were created and forecasting was done for the number of persons killed for the year 2006. The actual figures were compared with predicted figures during the forecasting window of 2001–06. Based on the results, ARIMA (7, 2, 0) has been used for estimation for the year 2006. Figure 7 shows the TS plot for forecasting till 2006. We can infer that the TS model performs reasonably well, supported statistically by high white noise and low RMSE value.

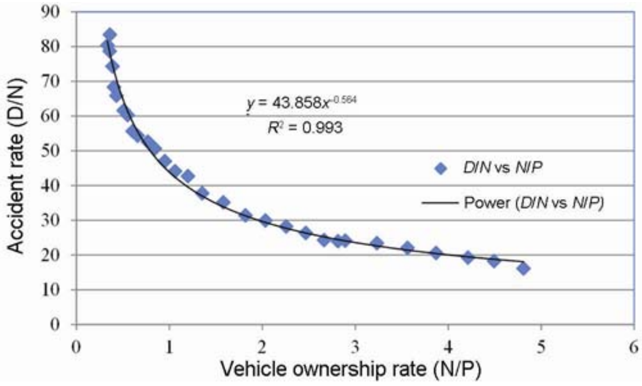

Figure 6. Best-fit curve for accident rate versus vehicle ownership rate.

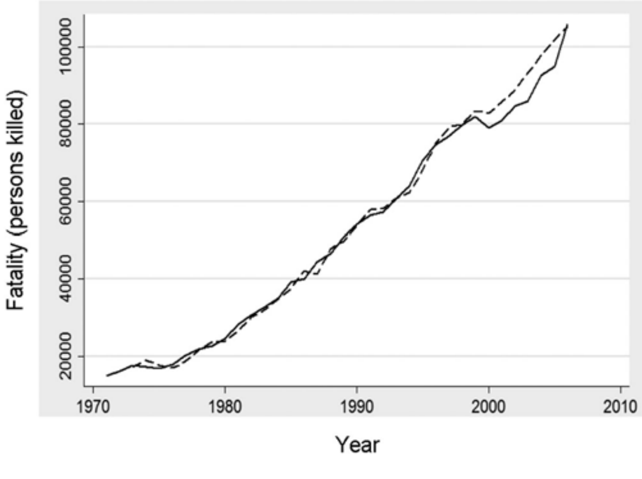

Figure 7. Actual and predicted TS plot of persons killed (D) till 2006.

In order to gauge the relative accuracies of the two methods explained above, the values for variables N, D and P were substituted in the regression equation, and D was calculated for the year 2006. This value of D was compared with the one obtained for 2006 from TS analysis of the same data. After comparison, the results from TS analysis were found to be more accurate and going a step further, forecasting was done for total number of persons killed for the year in 2021 using data from 1971 to 2006. ARIMA (2, 1, 0) was found out to be best suited for forecasting based on high white noise probability, and low AIC and BIC values.

Analysis with PeMS data: PeMS is an archived data user service that provides over ten years of data for historical analysis. In the raw AADT data available on the PeMS website, the column ‘Arithmetic Mean’ is the average of all daily flows. Each row shows this value for a year; so if the row starts at 4/1/2009 (in mm-dd-yyyy format), the value being shown is the arithmetic mean (the simple average) of daily traffic volumes from 4/1/2009 to 3/31/2010. The next row that starts at 5/1/2009 shows the arithmetic mean from 5/1/2009 to 4/30/2010 and so on. Study of these data for Lark Ellen (34.4 miles along I-10W) has been done. Table 2 gives the AADT data for this location. The choice of location is based on the following three different criteria: (i) The location should fall somewhere midway along the length of I-10, which itself is 46.8 miles long in District 7. (ii) Preferably, mainline data should be considered for analysis. (iii) That location should be selected for which data are available for the longest duration.

For this set of analysis, monthly AADT data from July 2000 to December 2008 were taken to estimate the AADT for March 2011 (27 data points ahead in future), the most recent point of time for which data were available. Table 7 shows important parameters for some prospective models. Almost all the models investigated have similar values for AIC and BIC, and thus we depend on high white noise and low RMSE for choosing the best model.

Table 8 shows statistics for ARIMA (2, 1, 2), which has been found to be best suited for modelling this case based on the results represented in Table 7. As can be noticed from the last two columns in Table 8, lag 1 for AR and both lags 1 and 2 for MA are significant at 5% level of significance, because the 95% confidence interval for these lags is far from zero. Portmanteau test for

Table 7. TS Parameters for prospective models for Lark Ellen

Model	P (white noise)	AIC	BIC	RMSE
ARIMA (1, 1, 1)	0.703	1523.31	1533.77	948.004
ARIMA (1, 1, 2)	0.847	1523.86	1536.94	953.260
ARIMA (2, 1, 2)	0.846	1523.05	1538.74	948.004
ARIMA (2, 1, 1)	0.825	1524.07	1537.15	956.340

Table 8. Regression parameters for ARIMA (2, 1, 2)

ARIMA regression					Number of obs = 101	
Sample: Sample: 2000m8–2008m12					Wald chi^2 (4) = 133.68	
Log likelihood = –755.5266					Prob > chi^2 = 0.0000	
D$_2$.TVP		OPG	*z*	*P* > |*z*|	(95% conf. interval)	
TVP	Coef.	Std. Err.				
_cons	–11.0944	126.6279	–0.09	0.930	–259.2806	237.0918
ARMA						
ar L1.	–0.5027332	0.215281	–2.34	0.020	–0.9246762	–0.0807903
L2.	0.2845574	0.216591	1.31	0.189	–0.1399531	0.7090679
ma L1.	1.308261	0.2071327	6.32	0.000	0.9022883	1.714234
L2.	0.4858782	0.1615852	3.01	0.003	0.169177	0.8025794
σ	427.2118	15.14778	28.20	0.000	397.5227	456.9009

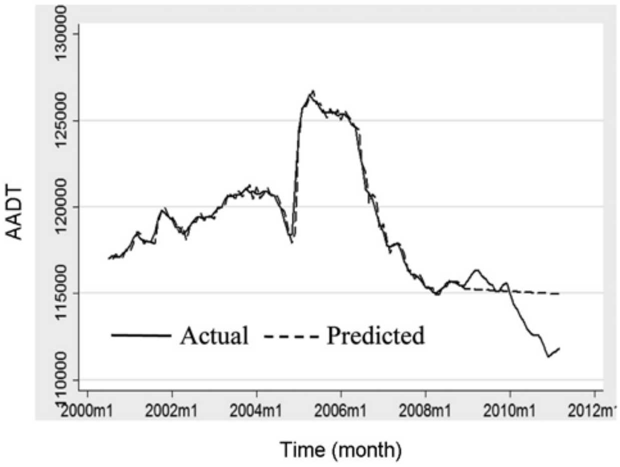

Figure 8. Actual and predicted AADT for Lark Ellen.

white noise gives Portmanteau (*Q*) statistic as 30.9877, which is less than the critical value of 55.76 at 5% level of significance. Probability >χ^2 (40) is 0.8459. Figure 8 shows the actual and predicted AADT values for Lark Ellen after analysis with ARIMA (2, 1, 2).

As done with the previous set of analysis, in order to investigate the effect of proximity of dataset to the forecasting window, analysis was done for Lark Ellen with the same dataset, but this time using data from July 2002 to December 2008 to approximate figure for March 2011, thus using only 78 data points for model-building as opposed to 102 in the previous case. For this analysis, ARIMA (2, 1, 2) was chosen out of six prospective models for forecasting.

Results

Table 9 gives the results of the analyses carried out using MoRTH data for varying (and increasing) number of years. The estimated and actual values are for the target year 2006. As discussed earlier, data used for 15 and 22 years could not be found sufficient enough for appropriate analysis. Results for other sets of data have been discussed subsequently.

A negative error signifies underestimation, i.e. the actual vehicular population is greater than the one predicted. A positive error shows that the value predicted by analysis is more than the actual value. As can be noted, TS analysis was possible only with data for 22 years and more (analysis with 15 and 20 years data could not be run due to data insufficiency). Further, the efficacy of TS analysis improved with increasing number of years. In case of predicting the fatality rate for the target year 2006 using two different approaches, the values obtained from regression and TS analysis were found to be 121,020 and 105,201 respectively, against the actual value of 105,700 for 2006, giving a percentage error of 14.49 and 0.47 respectively. Also, the average percentage error during the forecasting window (period 2001–06) was 17.67 for regression and 5.31 for TS analysis.

Moreover, it was observed that the error for forecasting for the year 2006 remained almost at the same if data till 2005 were used for model-building, irrespective of the amount of data used. This is substantiated by the results obtained from analyses done on data from 1981 to 2005 (25 years), 1976 to 2005 (30 years), and 1971 to 2005 (35 years), where their respective errors remained 2.24%, 2.44%, and 2.48%. It can also be observed that for the same amount of data, the dataset closer to the forecasting window gives better results compared to one with equal amount of data but farther from the forecasting window (for example, with 30 years dataset, the period 1976–2005 gives more accurate results than the period 1971–2000). This corroborates that accuracy of TS analysis improves with proximity to the forecasting window.

For the analysis carried out with AADT data from PeMS, DoT, California, ARIMA (2, 1, 2) predicts an AADT value of 114,959 for March 2011 while the actual value is 111,834, resulting in an overestimation error of 2.794%. In terms of numbers, these are noticeably lower than the ones obtained in the results for the Indian data. This highlights the fact that the efficiency of estimation by TS analysis improves drastically with increasing

Table 9. Results obtained with different sets of data from MoRTH

Data used	Model	Predicted figure for 2006 (in '000)	Actual figure for 2006 (in '000)	Absolute Error	Error (–ve; %)
1971–1996 (26 years)	ARIMA (3, 2, 0)	67,443	89,618	22,175	24.75
1971–2000 (30 years)	ARIMA (6, 2, 0)	73,152	89,618	16,466	18.37
1971–2005 (35 years)	ARIMA (1, 2, 0)	87,319	89,618	2299	2.48
1981–2005 (25 years)	ARIMA (1, 2, 0)	87,610	89,618	2008	2.24
1976–2005 (30 years)	ARIMA (1, 2, 0)	87,436	89,618	2182	2.44

amount (richness) of data available. Also, when analysis was done using data from July 2002 to December 2008 (thus using 24 data points lesser than the previous case) to predict for March 2011, ARIMA (2, 1, 2) resulted in a figure of 114,174, thereby showing an overestimation of 2.09%. Thus, as in the case of analysis with MoRTH data, the effect of proximity of sample data to the forecasting window seems to have a large bearing on the accuracy of TS forecasting, irrespective of the amount of data used.

For a similar data set sourced from the Indian Roads Congress (IRC), the error resulting from estimation of total vehicular population for the year 1996 using data for years 1951–85 (35 years) was 1.49%. The corresponding values of error obtained from the other two approaches – trend line analysis and econometric regression analysis were found to be 93.57% and 6.202% respectively[31]. Thus the error level of the results obtained from TS analysis is considerably lower than that from the other two methods, underlining its usefulness as a forecasting technique in the future.

Conclusion

Identification, investigation and implementation of appropriate traffic forecasting techniques are imperative to meaningful and sustainable allocation of scarce resources like land, labour and fund for developing nations. TS analysis can be a promising alternative to the problem of overestimation of future traffic levels, a trend generally observed when forecasting with other traditional techniques. The error in estimation for TS analysis was found lower in most of the cases considered. While the regression model for predicting fatality rate had a high R^2 value (0.993), the results from TS analysis showed better accuracy with lower average errors (5.31% and 17.67% for TS and regression analyses respectively). In a previous work, even for analysis done with IRC data using trend line and econometric approaches in addition to TS analysis, the error with TS modelling was considerably lower than the other two models (1.49% for TS, 6.202% for econometric and 93.57% for trend line analyses).

It can also be concluded that while as with other types of regression and statistical analyses, the more the quantity of data used (sample data), the better the analysis in terms of validating the models, their statistical coefficients and outputs, the results of TS analysis are heavily influenced by the proximity of the forecasting window to the sample data used for model building. Even with larger volume of data, the error levels for prediction remain constant enough as long as the forecasting window remains as close in all the cases. This tends to strongly suggest that TS models can be effective in short-term forecasting, underlining their potential use in fields that depend heavily on accurate short-term forecasts.

TS models have been found to be robust and work well with data for Indian as well as US locations. Although this approach requires rich data for variables used in modelling, it has the capability of accurate predictions using lesser number of explanatory variables than some other traditional approaches; thus the high data requirement for some variables is offset by requirement of lesser number of variables altogether. The time-frame for accurate forecasts using this method (which in this study is 15 years into the future) can be further investigated. As has been realized during this analysis, at least 30 data points are required for acceptable results from forecasting. If the limitation of high and rich data requirement for this method is overcome by implementation of proper technology then, in agreement with the findings of other researchers, it should contribute favourably towards accurate traffic forecasting (especially short-term forecasting) in the times to come.

It can be established that use of TS models can be effective in short-term forecasts in the range 5–10 data points ahead. This opens up vast avenues for its use in installation of facilities like ITS, traffic management and other congestion mitigation measures because these facilities majorly require short-term forecasts for traffic volume. Seasonal ARIMA (SARIMA) models are used in conjunction with Kalman filters for short-term traffic volume forecasting. *Inter alia* accurate forecasts for volume (flow) will result in better forecasts for speed, area occupancy and other useful derived parameters. Moreover, it is noticed that the TS method relies exclusively on historical trend and is not behavioural in nature, i.e. it does not take into account other explanatory variables that usually affect travel demand.

These models may also be used for predicting travel time and density in real-time traffic situations to predict the bottleneck traffic conditions and take traffic management measures, both for private and public transport. This

Appendix 1. Abbreviations used in the article

Abbreviation	Expanded form	Definition
AADT	Annual Average daily traffic	It is the average daily traffic for a year, i.e. the average daily traffic over a 365-day period.
ACF	Auto-correlation factor	It is the function representing correlation between the target variable and lag values for the same variable in a time series.
AIC	Akaike information criterion	For a given set of data, it is a measure of the relative quality of a statistical model; deals with the trade-off between the complexity and the goodness-of-fit of the model.
ARIMA	Auto-regressive integrated moving average	A variant of regression used to work with time-series data in statistics and econometrics.
BIC	Bayesian information criterion	Analogous to AIC, it resolves the problem of overfitting caused due to addition of parameters by introducing a penalty term for additional parameters used in the model.
CES	Consulting Engineering Services	–
DoT	Department of Transportation	–
GARCH	Generalized auto-regressive conditional heteroskedasticity	ARMA models that account for characteristic size or variance of error terms and volatility of error; model error terms as a function of size of error terms previous time periods.
GNP	Gross national product	It is the market value of all the products and services produced in one year by labour and property supplied by the residents of a country.
ITS	Intelligent transportation systems	Advanced engineering applications aimed at providing innovative services related to modes of transport and traffic management.
MoRTH	Ministry of Road Transport and Highways, Government of India	–
NHDP	National highways development project	A project to upgrade, rehabilitate and widen major highways in India to a higher standard; implemented in 1998 by National Highways Authority of India.
NSDP	Net state domestic product	Equals the gross domestic product minus depreciation on the capital goods of a state; accounts for capital that has been consumed over the year in the form of housing, vehicle or machinery deterioration.
OLS	Ordinary least squares	A method for estimating the unknown parameters in alinear regression model; minimizes the sum of squared vertical distances between the observed responses in the data set and the responses predicted by the linear approximation.
PACF	Partial auto-correlation factor	It is a function representing the autocorrelation of time-series observations separated by a lag of time units with the effects of the intervening observations eliminated.
RMSE	Root mean square error	A measure of the difference between values predicted by a model and those actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.
TS	Time series	A sequence of observations ordered with respect to time, typically at uniformly spaced time intervals.

may be possible with fusion of time-series data. On the other hand, regression models, while taking into account the dependency of travel demand on exogenous influencing variables, ignore the historical growth trend. A combination of these two approaches that makes use of exogenous explanatory variables in conjunction with seasonal/historical variations can be expected to result in better forecasts than either of these two approaches employed alone, especially with regard to long-term prediction. This warrants the use of multivariate TS methods like GARCH and ARCH processes, incorporating factors like change in land-use patterns and a few relevant economic indicators which should produce even more accurate results.

1. Dhingra, S. L. *et al.*, Application of time series techniques for forecasting truck traffic attracted by the Bombay metropolitan region. *J. Adv. Transp.*, 1993, **27**(3), 227–249.
2. Matas, A. *et al.*, Demand forecasting in the evaluation of projects. Working Paper in Economic Evaluation of Transportation Projects, Centro de Estudios y Experimentación de Obras Públicas (CEDEX), 2009, pp. 1–31.
3. Skamris, M. K. and Flyvbjerg, B., Inaccuracy of traffic forecasts and cost estimates on large transport projects. *Transp. Policy*, 1997, **4**(3), 141–146.
4. Cervero, R., Are induced traffic studies inducing bad investments? *ACCESS*, 2003, **22**, 22–27.
5. Cervero, R. and Hansen, M., Induced travel demand and induced road investment: a simultaneous equation analysis. *J. Transp. Econ. Policy*, 2002, **36**(3), 469–490.
6. Hymel, K. M. *et al.*, Induced demand and rebound effects in road transport. *Transp. Res. Board, Methodol.*, 2010, **44**(10), 1220–1241.
7. Ramsey, S., Of mice and elephants. *ITE J.*, 2005, **75**(9), 38.
8. Clark, S., Traffic prediction using multivariate nonparametric regression. *J. Transp. Eng., ASCE*, 2003, **129**(2), 161–168.
9. Kadiyali, L. R., Road transport demand forecast for 2000 AD. *J. Indian Roads Congress*, 1987, **48**(3), 353–432.

10. Kadiyali, L. R. and Shashikala, T. V., Road transport demand forecast for 2000 AD revisited and demand forecast for 2021. *J. Indian Roads Congress*, 2009, **557**, 235–237.
11. Project: Feasibility for 6-laning of NH-2 from Delhi–Agra project on DBFO pattern under NHDP Phase V, Consulting Engineering Services, New Delhi, India, October 2007, chapter 3.
12. Bhar, L. M. and Sharma, V. K., Time-series analysis. Indian Agricultural Statistics Research Institute, New Delhi, 2005, pp. 1–15.
13. Nihan, N. L. and Holmesland, K. O., Use of Box and Jenkins time series technique in traffic forecasting. *Transportation*, 1980, **9**(2), 125–143.
14. Oswald, R. K. *et al*., Traffic flow forecasting using approximate nearest neighbor nonparametric Regression. Research Report No. UVACTS-15-13-7, Center for Transportation Studies at the University of Virginia, USA, 2001.
15. Gazis, D. and Knapp, C., On-line estimation of traffic densities from time series of traffic and speed data. *Transp. Sci.*, 1971, **5**(3), 283–301.
16. Levin, M. and Tsao, Y., On forecasting freeway occupancies and volumes. *Transp. Res. Rec.*, 1980, **773**, 47–49.
17. Ahmed, M. S. and Cook, A. R., Analysis of freeway traffic time-series data by using Box–Jenkins techniques. *Transp. Res. Rec.*, 1982, **722**, 1–9.
18. Okutani, I. and Stephanedes, Y., Dynamic prediction of traffic volume through Kalman filtering theory. *Transp. Res., Part B*, 1984, **18**(1), 1–11.
19. Moorthy, C. K. and Ratcliffe, B. G., Short term traffic forecasting using time series methods. *Transp. Plann. Technol*., 1988, **12**(1), 45–56.
20. Stamatiadis, C. and Taylor, W., Travel time predictions for dynamic route guidance with a recursive adaptive algorithm. In Paper Presented at the 73rd Annual Meeting of Transportation Research Board, Washington, DC, USA, 1994.
21. Hamed, M. M., Al-Masaeid, H. R. and Said, Z. M. B., Short-term prediction of traffic volume in urban arterials. *J. Transp. Eng., ASCE*, 1995, **121**(3), 249–254.
22. Chang, J. L. and Miaou, S. P., Real-time prediction of traffic flows using dynamic generalized linear models. *Transp. Res. Rec.*, 1999, **1678**, 168–178.
23. D'Angelo, M., Al-Deek, H. and Wang, M., Travel time prediction for freeway corridors. *Transp. Res. Rec.*, 1999, **1676**, 184–191.
24. Lee, S. and Fambro, D., Application of the subset ARIMA model for short-term freeway traffic volume forecasting. *Transp. Res. Rec.*, 1999, **1678**, 179–188.
25. Williams, B. M., Multivariate vehicular traffic flow prediction: an evaluation of ARIMAX modeling. In Paper Presented at the 80th Annual Meeting of Transportation Research Board, Washington DC, USA, 2001.
26. Ishak, S. and Al-Deek, H., Performance evaluation of a short-term time-series traffic prediction model. *J. Transp. Eng., ASCE*, 2002, **128**(6), 490–498.
27. Tang, Y. F. and Lam, W. H. K., Annual average daily traffic forecasts in Hong Kong. *J. East Asia Soc. Transp. Stud*., 2001, **4**(3), 145–158.
28. Smith, B. L., Williams, B. M. and Oswald, R. K., Comparison of parametric and nonparametric models for traffic flow forecasting. *Transp. Res. Part C*, 2002, **10**(4), 303–321.
29. Tang, Y. F., Lam, W. H. K. and Pan, L. P., Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong. *J. Transp. Eng. ASCE*, 2003, **129**(3), 223–329.
30. Chandra, R. S. and Al-Deek, H., Cross correlation analysis and multivariate prediction of spatial time series of freeway traffic speeds. *Transp. Res. Rec*., 2008, **2061**, 64–76.
31. Jha, K. *et al*., Modeling growth trend and forecasting techniques for vehicular population in India. *Int. J. Traffic Transp. Eng.*, 2013, **3**(2), 139–158.
32. Box, G. E. P. and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*, 1976, Holden-Day, San Francisco.
33. Performance Measurement System Database, California Department of Transportation, USA; http://www.pems.dot.ca.gov (accessed on 29 March 2012)
34. Pankratz, A., *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, 1983, John Wiley, New York.
35. Naylor, T. H., Seaks, T. G. and Wichevn, D. W., Box–Jenkins methods: an alternative to economic forecasting. *Int. Stat. Rev*., 1972, **40**(2), 123–137.
36. Nelson, C. R., *Applied Time Series Analysis: For Managerial Forecasting*, Holden-Day, San Francisco, 1973, pp. 139–169.
37. Vishwas, M. *et al*., Some issues pertaining to sustainability of road transport operations, road construction and maintenance in India over the next twenty years. *J. Indian Roads Cong.*, 2012, **73**(2), 135–158.

ACKNOWLEDGEMENTS. We thank Jane Berner (Caltrans) for providing access to the PeMS data and for also providing useful suggestions and illustrations about the database, and N. K. Jha (O. P. Jindal School, Patratu) and A. K. Giri (Birla Institute of Technology & Science (BITS), Pilani) for their useful insights and contributions.

Received 12 September 2014; revised accepted 19 February 2015

doi: 10.18520/cs/v110/i3/373-385