International Journal for Traffic and Transport Engineering, 2013, 3(2): 139 – 158
DOI: http://dx.doi.org/10.7708/ijtte.2013.3(2).04 UDC: 330.43:519.246(540)

MODELING GROWTH TREND AND FORECASTING TECHNIQUES FOR VEHICULAR POPULATION IN INDIA

Kartikeya Jha[1], Nishita Sinha[2], Shriniwas Shrikant Arkatkar[3], Ashoke Kumar Sarkar[4]
[1, 3, 4] *Department of Civil Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India - 333031*
[2] *CITD, Jawaharlal Nehru University, New Delhi, India - 110067*

Received 4 January 2013; accepted 21 March 2013

Abstract: Forecasting and estimation of growth in vehicular population is a sine qua non of any major transportation engineering development, requires capturing the past trend and using it to predict the future trend based on qualified assumptions, simulations and models created using explanatory variables. This work attempts to review the in vogue approaches and investigate a more contemporary approach, the Time Series (TS) Analysis. Three fundamentally different methods were explored and results from each of these analyses were collated to check for respective levels of accuracy in predicting vehicular population for the same target year. Within the scope of this study and estimation, results obtained from TS Analysis were found to be considerably more accurate than those from Trend Line Analysis and observably better than those from Econometric Analysis. To reinforce these observations and inferences drawn, a second set of analysis was done on more recent input by using AADT data from PeMS, California. Inter alia this was carried out to contrast any statistical improvement observed when doing TS analysis with rich and accurate data. With all the data sets used and locations analyzed for forecasting, the Time Series analysis technique was invariably found to be a potent tool for forecasting.

Keywords: econometrics, time series, ARIMA, regression, elasticity, root mean square error.

1. Introduction

In a broad sense, traffic forecasting is the process of estimating the number of vehicles or people that are 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, the number of ships arriving at/leaving from a seaport or may estimate the expected future traffic levels for the whole country. This process begins with the collection of data on current traffic. This traffic data is 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 segment or railway station that falls under the scope of facility. Due to data availability constraints for the Indian case, the present analysis has been done for total vehicular population in India to develop an insight into the choice of appropriate methods for estimation at specific project level also.

[3] Corresponding author: sarkatkar@pilani.bits-pilani.ac.in

1.1. 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 (Dhingra et al., 1993). It is an essential input to start the planning and/or development phase of any major transportation project initiatives. It is the first step in defining the scope and geometry of such projects. In case of highways, the geometric and structural designs are based on forecast traffic volumes and the ESAL (Equivalent Single Axle Load) forecast. It is also used to assess the Level of Service (LoS) for the present and future traffic conditions. Sometimes forecasting even helps us know whether a project is needed at all. Forecasting is necessary for doing relevant economic analysis (Matas et al., 2009). 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 (Skamris and Flyvbjerg, 1997). The costs associated with an under designed project arise when an additional project must satisfy the original inadequacies. Extra materials, labor and additional right-of-way attainment add to the cost of an over designed project (IRC: SP-30, 1984). Usually, an estimation of traffic for 15 or 20 years after the construction date is considered. In India, National Highways are designed for 15 years after completion of the work. 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 the future (Moeinaddini et al., 2012), especially Induced Traffic Effect (Cervero and Hansen, 2002; Cervero, 2003) which says that the improvement in transportation facilities more often than not results in generation of new traffic demand because of new users who did not use these facilities when the new developments were not in place i.e. increase in highway capacity attracts new traffic, and the Rebound Effect (Hymel et al., 2010) which highlights the more-than-before use of transportation facilities by the current users due to availability of new and improved services. Both of these effects result in the complete benefits of new transportation facilities not reaching the users as a result of increased use of existing users or an increase in the number of end users,

- Traffic impacts due to development in infrastructure, majorly due to change in land use patterns (Ramsey, 2005) which may severely affect the estimated demand for traffic in the future. This may result in underestimation (if there are major industrial developments in that corridor in near future) or overestimation (if there are major industrial shifts away from that area),
- Unforeseen and unaccounted socio-economic changes (Clark, 2003), and
- Construction of new roads, diversions etc.

1.2. Challenges Faced

The following are the major challenges confronted during the process of forecasting future traffic volumes:

- Due to construction or improvement of physical facilities – which affect the cost and quality of transport services, the traffic forecasts may overestimate or underestimate the actual figures.
- Unaccounted induced demand effect and rebound effect may lead to serious

errors in estimation and forecasting (Parthasarathi, 2001).

- This invokes the need to understand and predict (to some extent) the effect of incorporating new traffic and/or road facilities on human behavior.
- Traffic projection should reflect expected economic, demographic and land use trends, based on historic and projected relationships between these factors and regional traffic growth.
- Some studies show a tendency for forecasts of construction costs to be underestimated, and for traffic forecasts to be overestimated.
- A review conducted by Standard and Poor's (Bain and Plantagie, 2004), with its suggested mean of 76%, tends to suggest that, on average, toll road traffic forecasts overestimate year one traffic by 20% - 30%. The mean value of 76% shows that the actual traffic was only 0.76 times that predicted (forecast).
- Forecasting may turn out to be a difficult assignment, especially when done for short term intervals. This is majorly because unforeseen exigencies are very difficult to predict and account for. The capacity of the road system can be reduced by exogenous factors like demonstrations, roadwork or the weather.

2. 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 Analysis. In the Indian context, the past research work has mainly concentrated on Trend Line analysis (Kadiyali, 1987; Kadiyali and Shashikala, 2009). Here the traffic volume levels for the country have been predicted using a linear relationship between the country's Gross National Product (GNP) and the total vehicular population data. On the same lines, a project feasibility report on 6-laning of NH-2 (Feasibility for 6-laning of NH-2 from Delhi-Agra Project on DBFO pattern under NHDP Phase V, Consulting Engineering Services, Oct 2007) elaborates a combination of Trip Generation models and Trend Line Analysis using NSDP (Net State Domestic Product) instead of GNP for different corridors lying in scope of the project. Study of more contemporary areas of research focuses mainly on the Time Series analysis. While Time Series Analysis (Bhar and Sharma, 2005), IASRI, New Delhi deals with the applications and nuances of Time Series analysis (exemplified with the use of the software SPSS), Nihan and Holmesland (1980) stress on the basics of Time Series Modeling. Approximate nearest neighbor nonparametric regression method has been discussed by Oswald et al. (2001). Traffic forecasting is a process predicting a dynamic variable. That is why a number of approaches may be adopted for traffic forecasting depending upon the situation at hand. Although there can be various methods for traffic volume forecasting, for this analysis the three most relevant methods were chosen for a comparative analysis due to data availability constraints.

The practice hitherto adopted in India (Nanda, 2005) is to collect seven days traffic volume counts twice a year on National Highways and 1 to 3 days on other roads. The NHDP programs generally follow three techniques for traffic projection:

- Past Trend Data: Available records over the past 5-10 years are collected from NH Division or respective PWD departments. This data is the base for observing the

traffic growth of the present scenario and the same trend is used to predict the future demand. This method is adopted at feasibility assessment stage.

- Vehicles Registration: Vehicle registration data is obtained from Motor Transport Statistics of India. Based on this data it is possible to predict the growth rate of vehicle registration of each individual vehicle type through regression analysis. This technique too is adopted at feasibility stage/level.
- Elasticity of Transport Demand: The projected future traffic incorporates analysis of some of the key socio-economic characteristics and rate of change expected during the study period in the project influence area. In India, the computed elasticities are comparable to the World Bank's recommended elasticities. This method is adopted at detailed project preparation stage.

3. Methodology of Analysis

This work attempts to offer a review of the eminent issues that confront traffic forecasting with the aim to avoid major errors that might distort the final outcome of evaluation. An attempt has been made to do an analysis of different traffic forecasting techniques by using input data from a common source and comparing the results obtained after forecasting with actual figures for each such method/technique used. In this process, three techniques have been adopted for comparison of results – Trend Line Analysis, Econometric Analysis and Time Series Analysis.

3.1. Approach and Data Collection

This whole exercise is only the first step in developing an insight into the choice of the best suited method, especially with respect to Indian conditions to estimate future traffic levels in the country which, as has been discussed, is quite imperative from many aspects. Due to data availability constraints this analysis has been done for total vehicular population in India. The primary data used has been cited from "Time Series Data on Road Transport Passenger and Freight Movement (1951-1991)", Special Publication 45, Indian Roads Congress, New Delhi, 1996. A part of it has been reproduced in Table 1 for ready reference. Since such analysis has been done which involves the comparison of results obtained from three different methods, a time period was selected for which data for all the input variables were available for all these three methods for the corresponding years.

To highlight the efficacy of the Time Series Analysis in forecasting, especially short-term forecasting, a separate set of analysis on the AADT data taken from PeMS (Performance Measurement System), Dept. of Transportation, California, US has been done for a location on the Interstate 10 (running from West to East). PeMS is an Archived Data User Service (ADUS) that provides over ten years of data for historical analysis. For data analysis, AADT data for location Lark Ellen (34.4 miles from west) along the Interstate 10, California falling in District 7 and running from West to East direction has been taken. The data from this station were found to be more consistent as per the detector health plots (plot that signify consistency and accuracy of raw data collected by sensors) obtained from PeMS. The corresponding data used for analysis is reproduced in Table 2.

3.2. Methods Adopted

Three fundamentally different methods have been chosen for comparative analysis of the results. These are briefly described below:

- Trend Line Analysis: This assumes a linear relationship between country's Gross National Product (GNP) and the total vehicular population. Generally, an equation of the form given below (Eq. (1)) is applied to arrive at desired results:

$$Log\,(T) = a + b * \log(GNP) \qquad (1)$$

Where,

T = Transport Demand
GNP = Gross National Product
a and *b* are coefficients which are derived empirically.

- Econometric Analysis: The traffic growth is seen as being dependent on certain economic and demographic indicators (Gujarati, 2004) such as Population, Per Capita Income/Per Capita Net National Product etc. Usually this method lends more logic and dependability to the estimation process. Many different combinations of economic/demographic indicators such as population, Per Capita Income (PCI)/Per Capita Net National Product (PCNNP), total labor force (urban & rural), total employed population (urban & rural) etc. can be tried to bring out the best results (Baltagi, 1999) but in this case, due to data availability constraints, a combination of two significant variables, namely Population and Per Capita Income has been taken for analysis.
- 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 suggested by Box and Jenkins (1976), ideally at least 50 observations are required for performing appropriate Time Series Analysis. The Box and Jenkins methodology (Pankratz, 1983) has been adopted and analysis has been done using the Auto-Regressive Integrated Moving Average (ARIMA) approach. The main reason behind using Box and Jenkins technique is that it has been shown to give relatively accurate forecasts. The results from comparative studies conducted by Naylor et al. (1972) and Nelson (1973) show that the Box and Jenkins model, although simpler, was more effective than other such contemporary econometric models.

4. Data Analysis

The complete data analysis is segmented into two sections. The first set of analysis deals with the data obtained from IRC for which analysis by all the three methods has been performed. For the second set, Time Series Analysis has been performed on data sourced from PeMS, DOT, California (CALTRANS). All these sets of analyses have been presented sequentially hereafter.

4.1. Analysis with IRC Data

For the first case, the data used has been cited from "Time Series Data on Road Transport Passenger and Freight Movement (1951-1991)", Special Publication 45, Indian Roads Congress, New Delhi, 1996. Table 1 gives the relevant data which have been used for analysis.

On an average, all vehicle categories as well as the total vehicular population appear to follow an exponential growth pattern, the slope of which seems to increase after year 1988-89 for most of the categories. The method-wise procedure for analysis using all the three methods has been briefly elaborated along with relevant illustrations.

4.1.1. Trend Line Analysis

The IRC data for vehicular population has been used to establish the relation "Log T= a + b Log GNP". The data used for analysis is for the years 1961-1985 (25 years) and estimation has been done for the year 1996 (11 years ahead in future). The plot and the corresponding relation between Log T (T = total vehicular population) and Log GNP (Eq. (2)) can be seen in Fig. 1.

The equation that emerged after linear regression is:

$$Log\,(T) = 2.695 * Log\,(GNP) - 7.708 \qquad [R^2 = 0.989] \qquad (2)$$

4.1.2. Econometric Analysis

The economic/demographic indicators chosen for analysis are Population and Per Capita Income. The Employment Rate and Total Available Labor Force figures could not be used for analysis due to lack of data (insufficient number of data points) as only decennial data was available for these variables (Green, 1993). Moreover, the choice of the two indicators seems justifiable since this captures the effect of both significant factors - the number of users in the present and future (population) as well as the purchasing power of these users (Per Capita Income). Here also the data for 1961-1985 has been taken for model building and estimation has been done for the year 1996. The available data were subjected to Regression Analysis on SPSS (Multiple Linear Regression). The analysis has been done for the same period as that for Trend Line Analysis. Eq. (3) comes out as the result of the regression analysis:

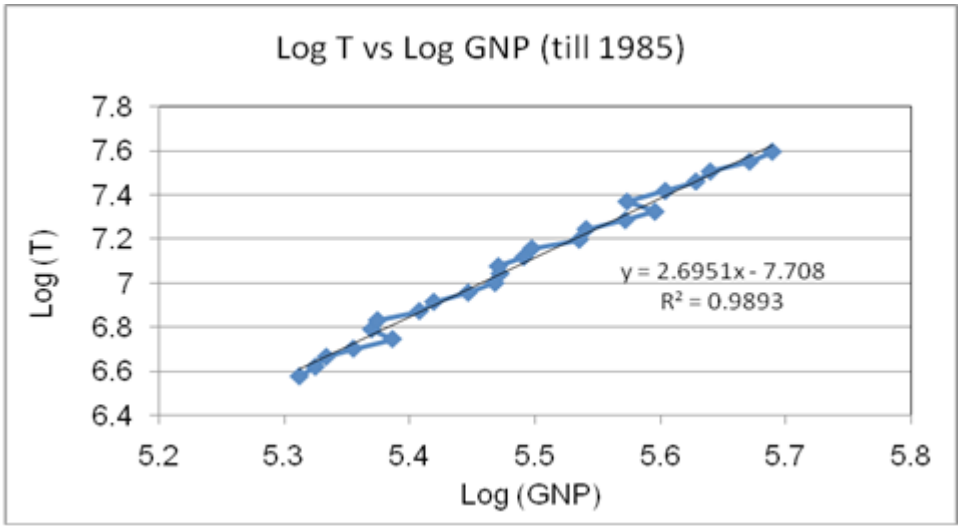


Fig. 1.
Plot between Log T & Log GNP

Table 1

Data Table for Variables Used in Analysis

Year	Vehicular Population	GNP (Rs. Crores)	Per Capita Income (Rs.)	Population	Year	Vehicular Population	GNP (Rs. Crores)	Per Capita Income (Rs.)	Population
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1961	3778488	205196	1350	439435	1979	21083477	394335	1689	658730
1962	4173044	211287	1355	452378	1980	23418452	374640	1550	688956
1963	4630391	215601	1347	462196	1981	26138616	401970	1630	683329
1964	5042291	226577	1384	472305	1982	28846935	425168	1693	703570
1965	5574485	243472	1454	482706	1983	32056201	436577	1691	719090
1966	6172690	234394	1355	493389	1984	35530913	469293	1790	734870
1967	6786859	236846	1335	504345	1985	39429002	489206	1811	749184
1968	7466313	255843	1413	515601	1986	38349721	511058	1841	767200
1969	8219423	262687	1415	527177	1987	45492645	532021	1871	783730
1970	9049346	279791	1478	539075	1988	53073160	551409	1901	797526
1971	10014079	293933	1520	548160	1989	60827580	607207	2059	817490
1972	11028301	296688	1492	563530	1990	68944375	648108	2157	833929
1973	11918799	295752	1446	575887	1991	74641916	683670	2222	846421
1974	13109888	309950	1483	588299	1992	80487495	691143	2175	883473
1975	14359564	314509	1469	600763	1993	86298645	726375	2239	900453
1976	15717431	343173	1572	613273	1994	92274138	769265	2292	918570
1977	17553280	347530	1552	630200	1995	100337963	824816	2449	934228
1978	19303907	373464	1635	644330	1996	108336195	886961	2573	945121

Column (2): Source -"Time Series Data on Road Transport Passenger and Freight Movement (1951-1991)", Special Publication 45, IRC, New Delhi, 1996

Column (3): Source - Economic Survey, 2007-08, Govt. of India, Oxford University Press, New Delhi, 2008 (GNP at 1993-94 prices)

Column (4) & (5): Source - India: Economic Indicators (1951-2001), Website (http://www.indiastat.com) accessed on 21.02.2012 (Per Capita Income at 1980-81 prices)

Table 2

Monthly AADT Data for Location 'LARK ELLEN'

Month	Mainline (ML) AADT	Month	Mainline (ML) AADT	Month	Mainline (ML) AADT
Jul-00	117018	Feb-04	120744	Sep-07	116075
Aug-00	117170	Mar-04	120914	Oct-07	116076
Sep-00	117113	Apr-04	120896	Nov-07	115953
Oct-00	117099	May-04	120710	Dec-07	115650
Nov-00	117339	Jun-04	120672	Jan-08	115364
Dec-00	117462	Jul-04	120258	Feb-08	115341
Jan-01	117725	Aug-04	119920	Mar-08	115093
Feb-01	118230	Sep-04	119160	Apr-08	115009
Mar-01	118441	Oct-04	118394	May-08	115176
Apr-01	118147	Nov-04	118376	Jun-08	115351
May-01	118070	Dec-04	121533	Jul-08	115630
Jun-01	118020	Jan-05	124428	Aug-08	115670
Jul-01	117952	Feb-05	125687	Sep-08	115698
Aug-01	118524	Mar-05	125992	Oct-08	115596
Sep-01	119371	Apr-05	126534	Nov-08	115441
Oct-01	119828	May-05	126209	Dec-08	115472
Nov-01	119624	Jun-05	126098	Jan-09	115743
Dec-01	119469	Jul-05	125727	Feb-09	116064
Jan-02	119249	Aug-05	125454	Mar-09	116292
Feb-02	118846	Sep-05	125454	Apr-09	116346
Mar-02	118790	Oct-05	125431	May-09	116031
Apr-02	118425	Nov-05	125496	Jun-09	115870
May-02	118615	Dec-05	125283	Jul-09	115563
Jun-02	118910	Jan-06	125427	Aug-09	115462
Jul-02	119169	Feb-06	125349	Sep-09	115122
Aug-02	119426	Mar-06	125197	Oct-09	115156
Sep-02	119412	Apr-06	124803	Nov-09	115486
Oct-02	119364	May-06	124623	Dec-09	115616
Nov-02	119450	Jun-06	123341	Jan-10	114961
Dec-02	119397	Jul-06	122467	Feb-10	114377
Jan-03	119689	Aug-06	121037	Mar-10	114107
Feb-03	119904	Sep-06	120781	Apr-10	113668
Mar-03	120019	Oct-06	120445	May-10	113356
Apr-03	120451	Nov-06	119571	Jun-10	112959
May-03	120665	Dec-06	118978	Jul-10	112688
Jun-03	120594	Jan-07	118558	Aug-10	112579
Jul-03	120750	Feb-07	117770	Sep-10	112617
Aug-03	120717	Mar-07	117690	Oct-10	112282
Sep-03	120831	Apr-07	117817	Nov-10	111810
Oct-03	121107	May-07	117860	Dec-10	111340
Nov-03	120832	Jun-07	117484	Jan-11	111574
Dec-03	120931	Jul-07	116822	Feb-11	111668
Jan-04	120731	Aug-07	116424	Mar-11	111834

Source: http://www.pems.dot.ca.gov (PeMS, Caltrans) accessed on 02.04.2012

$$Log\,(T) = -5.613 + 4.121 * Log\,(P) + 0.414 * Log\,(PCI) \quad (3)$$

Where,
T = Total vehicular population;
P = Total population;
PCI = Per Capita Income.

It can be noticed that the effect of Population on the total traffic demand is more than that of Per Capita Income. This seems to be logical to assume that to a certain degree, both these variables affect the traffic demand but this demand depends more on the number of prospective users than other factors such as their income levels. This may also be because vehicle categories like cycles, auto rickshaws, buses etc. have been considered whose number is not completely dependent on the income of the user only but to a large extent on their number. This is reflected in a lower value of elasticity for PCI (0.414) than that for Population (4.121). Table 3 and Table 4 show some important indicators which have been briefly explained subsequently. The Durbin-Watson test statistic tests the null hypothesis that the residuals from an ordinary least-squares regression are not auto-correlated against the alternative that the residuals follow an AR1 process. The Durbin-Watson statistic ranges in value from 0 to 4. A value near 2 indicates non-autocorrelation; a value toward 0 indicates positive autocorrelation while a value toward 4 indicates negative autocorrelation. Because

Table 3
Analysis Model Summary
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	0.999[a]	0.999	0.999	*****	0.999	8975.913	2	22	0.000	1.657

[a] *Predictors: (Constant), LogP, Log PCI*
[b] *Dependent Variable: LogT*

Table 4
Coefficient Statistics (IRC Data)
Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1 (Constant)	-5.613	0.293		-19.13	0.000					
LogPCI	0.414	0.161	0.053	2.58	0.017	0.939	0.482	0.019	0.130	7.68
LogP	4.121	0.090	0.949	45.92	0.000	0.999	0.995	0.343	0.130	7.68

[a] *Dependent Variable: LogT*

of the dependence of any computed Durbin-Watson value on the associated data matrix, exact critical values of the Durbin-Watson statistic are not tabulated for all possible cases. Instead, Durbin and Watson established upper and lower bounds for the critical values. Typically, tabulated bounds are used to test the hypothesis of zero autocorrelation against the alternative of positive first-order autocorrelation, since positive autocorrelation is seen much more frequently in practice than negative autocorrelation. The value of this statistic in our case is 1.657 which lies outside the limits dL = 1.21 and dU = 1.550 (for n = 25, n being the sample size; and 5% level of significance). If the observed value of the test statistic is less than the tabulated lower bound, then we should reject the null hypothesis of non-autocorrelated errors in favor of the hypothesis of positive first-order autocorrelation. If the test statistic value were greater than dU, we would not reject the null hypothesis. Since 1.657 > dU, we infer that the errors in this case are not auto-correlated. The t statistic and its significance value are used to test the null hypothesis that the regression coefficient is zero (or that there is no linear relationship between the dependent and independent variable). The 't-test' may be used to determine whether an estimated regression coefficient is significant, by forming the following ratio: t = the regression coefficient divided by standard error of the regression coefficient. Since the values of 't' for both these variables is greater than the critical value of 1.711 (for 24 degrees of freedom), these variables were found to be significant based on the results of the t-test. This is also reflected in their significance values of 0.017 for PCI and 0.000 for Population (both being less than 0.05). Table 5 shows important parameters for regression and residuals of the model.

The output for regression displays information about the variation accounted for by the model while that for residual displays information about the variation that is not accounted for by the model. A model with a large regression sum of squares in comparison to the residual sum of squares indicates that the model accounts for most of variation in the dependent variable. The F statistic is the regression mean square (MSR) divided by the residual mean square (MSE). If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable. Table 6 shows coefficients obtained for correlation and covariance.

The sign of the correlation coefficient indicates the direction of the relationship (positive or negative) while the absolute value of the correlation coefficient (which varies from -1 to 1) indicates the strength, with larger absolute values indicating stronger relationships. In the covariance matrices, the variances are displayed on the main diagonal and covariances are displayed above and below the main diagonal. The histogram (Fig. 2)

Table 5

Parameters for Regression and Residuals (IRC Data)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.286	2	1.143	8975.913	0.000[a]
	Residual	0.003	22	0.000		
	Total	2.289	24			

[a] *Predictors: (Constant), LogP, LogPCI*

[b] *Dependent Variable: LogT*

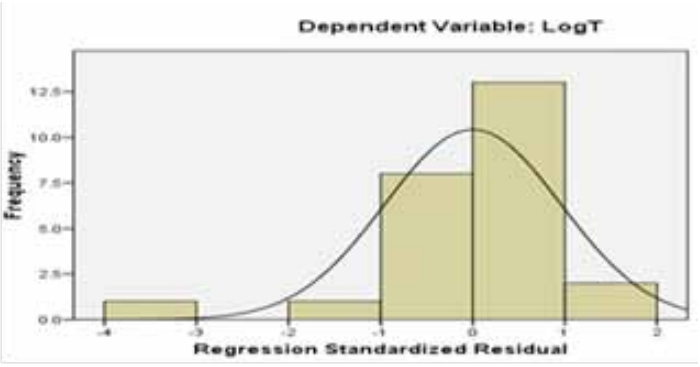
Table 6

Coefficient Correlations and Covariances (IRC Data)

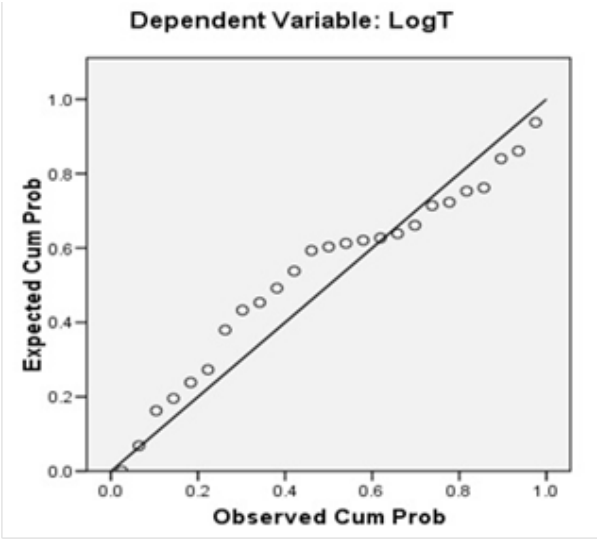
Coefficient Correlations[a]

Model			LogP	LogPCI
1	Correlations	LogP	1.000	-0.933
		LogPCI	-0.933	1.000
	Covariances	LogP	0.008	-0.013
		LogPCI	-0.013	0.026

[a] *Dependent Variable: LogT*



(a)



(b)

Fig. 2.

(a) Residual Histogram

(b) Normal P-P Plot of Regression Standardized Residual

is a frequency plot obtained by placing the data in regularly spaced cells and plotting each cell frequency versus the center of the cell. This graph is used to verify that the residuals are normally distributed, as is assumed by the regression model. Due to lesser number of observations, a perfect normal graph has not been obtained here but the plot seems to suggest normal distribution of residuals, and hence the error terms can be said to be normally distributed.

The basic idea behind any normal probability plot is that if the data follow a normal distribution with mean μ and variance σ^2, then a plot of the theoretical percentiles of the normal distribution versus the observed sample percentiles should be approximately linear. Since the concern is about the normality of the error terms, a normal probability plot of the residuals is created. If the resulting plot is approximately linear, the error terms are said to be normally distributed. This has been shown in Fig. 2.

For a better understanding of the situation, regression analysis was done using absolute figures (in place of logarithms) for PCI and Population. Table 7 shows some significant statistics for this analysis.

In the above analysis, the population has been taken in Millions. This indicates that if the population increases by a million, PCI remaining the same, the vehicular population should approximately increase by 71028 units. On the other hand, this figure increases by 25429 if the PCI increases by 1 Rupee, the population being constant. These figures appear to be reasonable in case of India. In this case also, the t-test confirms significance of variables.

4.1.3. Time Series Analysis

For univariate time series analysis, same data (from IRC) for the period 1951-1985 (35 years) has been used for analysis. The estimation has been done for the same target year 1996. The Box and Jenkins methodology has been used and ARIMA (Auto-Regressive Integrated Moving Average) technique has been adopted for analysis. The modeling has been performed on STATA (Hansen, 2007). The following brief definitions will enable a better understanding of the Time Series 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 data set such that assumption of stationarity is reasonable: Before identification of the model a basic

Table 7
Regression with Absolute Figures
Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations		
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part
1 (Constant)	-6E+007	7692640		-8.423	0.000	-80751607	-48844487			
AllIndpop	71027.9	14885.18	0.646	4.772	0.000	40158	101898	0.968	0.71	0.22
PCIRs	25429.3	10034.71	0.343	2.534	0.019	4619	46240	0.950	0.48	0.12

step of data transformation needs to be done where the variance of the data set is stabilised and the mean becomes constant. Dickey Fuller and Philip Perron Tests are performed to confirm stationarity of data used.

- Identification of the parameters of the model: ARIMA stands for Autoregressive Integrated Moving Average process. ARIMA modelling is in the very heart of univariate time series analysis. To get the order of AR and MA process, autocorrelation function and partial autocorrelation function are studied. An autoregressive process is a function of lagged dependent variables and a moving average process a function of lagged error terms. If a series needs to be differenced d times before it is stationary, I equals d for that series.
- Estimation of the parameters: There are two different ways a model can be estimated – *Maximum Likelihood Estimation* and *Conditional Maximum Likelihood Estimation*. The first one uses numerical optimization techniques for estimation purpose and the latter is simple OLS regression. This analysis follows full Maximum Likelihood Estimation. Based on AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) which determine the parsimony of the model, the 2 best models are ARIMA (3,2,3) and ARIMA

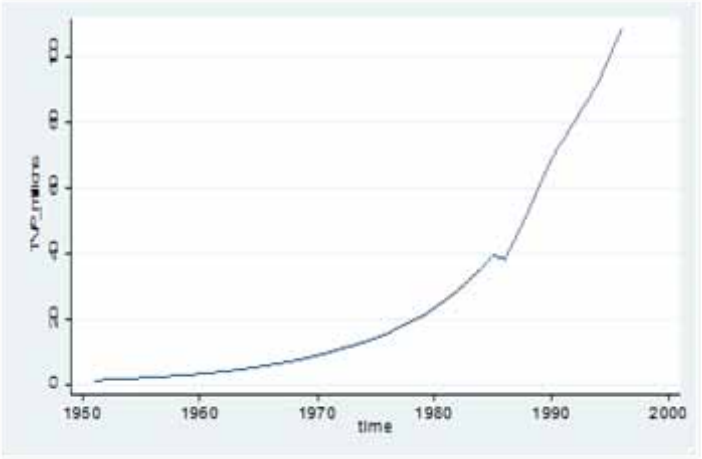
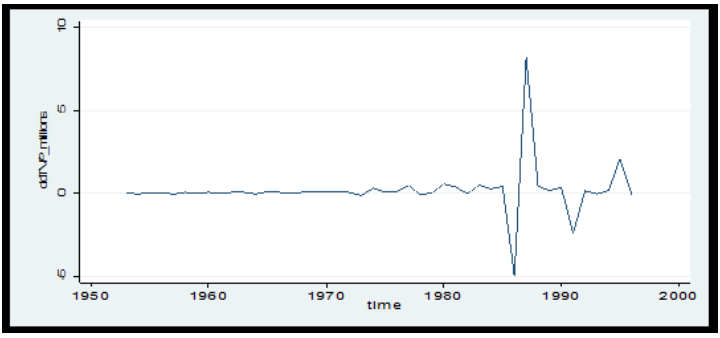


Fig. 3.
(a) Non-Stationary Raw Data



(b) Stationary Data after Double-Differencing

(7,2,3) for this case.

- 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*. Results are considered better when the value of this statistic is closer to 1.
- Forecasting: There are two kinds of forecasting that can be done using an ARIMA model – a static forecasting and a dynamic forecasting. The static forecast or the simple one step ahead forecast will forecast only for a single time period ahead at a time. Dynamic forecast on the other hand is used for forecasting for a longer horizon. The predicted or the forecast values will be the same as the one forecast by static forecast till the first extension of time period. For the purpose of forecasting, the period from 1986-1996 has been kept aside taken as the 'forecasting window'. These observed values will then be compared with the forecast values to calculate the root mean forecasting error.

The available data had to be differenced twice to achieve stationarity (a pre-requisite for Time Series Analysis) (Fig. 3). The Dickey Fuller and Philip Perron tests were conducted to confirm stationarity.

Out of 9 prospective models, the model ARIMA (7,2,3) was chosen based on least RMSE (Root Mean Square Error) and Maximum Likelihood Rule. The reason for this choice is derivable from Table 8.

Table 9 and Table 10 show component statistics for ARIMA (3,2,3) and ARIMA (7,2,3) respectively.

Table 12 shows statistics for ARIMA (2,1,2) which has been found to be best suited for modeling this data based on the results of these tests.

Fig. 4 shows the actual and predicted AADT values after analysis with ARIMA (2,1,2).

4.2. Time Series Analysis of Data from PeMS, DOT, California

This analysis uses AADT data extracted from PeMS, Dept. of Transportation, California, US. In the raw AADT data available on PeMS website the column that says "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. The data for this location have been more consistent (Table 2). That is why a reasonable level of accuracy was achieved even when data from Jul 2000 to Dec. 2008 were taken to estimate the AADT for Mar. 2011 (27 data points ahead in future). Important parameters for some prospective models have been shown in Table 11.

Thus in this case, a reasonable degree of accuracy has been achieved through consistency of data. The root mean square error values are also low as compared to those for the analysis with Indian data.

5. Results and Discussion

The results of the analyses carried out by different methods and the inferences drawn have been discussed here. The target year for estimation in the first set of analysis which uses data from IRC, New Delhi (involving all

Table 8

TS Test Statistics for Various Prospective Models

Model	P (White Noise)	AIC	BIC	RMSE
ARIMA(3,2,3)	0.932	882.4	894.4	2572936
ARIMA(4,2,3)	0.832	889.6	903.1	3674235
ARIMA(5,2,3)	0.906	885.1	900.0	1603122
ARIMA(6,2,3)	0.909	886.8	903.3	1593738
ARIMA(7,2,3)	0.998	882.6	900.6	1374773

Table 9

ARIMA (3,2,3)

ARIMA regression

Sample: **1953 – 1985**
Log likelihood = **-433.2066**

Number of obs = **33**
Wald chi2 (**6**) = **207.16**
Prob > chi2 = **0.0000**

D2.TVP	Coef.	OPG Std. Err.	z	P> |z|	[95% Conf. Interval]	
TVP						
_cons	164694.3	212519.4	0.77	0.438	-251836	581224.6
ARMA						
ar						
L1.	1.013409	1.03928	0.98	0.330	-1.023543	3.050361
L2.	-0.3121788	1.052095	-0.30	0.767	-2.374247	1.749889
L3.	0.267654	0.1870239	1.43	0.152	-0.0989062	0.6342141
ma						
L1.	-628.3457	519109	-0.00	0.999	-1018063	1016807
L2.	1009.124	946535.1	0.00	0.999	-1854166	1856184
L3.	-626.7141	693119.3	-0.00	0.999	-1357862	1357862
/sigma	-172.3563	161780.2	-0.00	0.999	-317255.7	316911

Table 10

ARIMA (7,2,3)

ARIMA regression

Sample: **1953 – 1985**
Log likelihood = **-429.2957**

Number of obs = **33**
Wald chi2 (**10**) = **595.10**
Prob > chi2 = **0.0000**

D2.TVP	Coef.	OPG Std. Err.	z	P> |z|	[95% Conf. Interval]	
TVP						
_cons	237106.3	268903.5	0.88	0.378	-289934.9	764147.5
ARMA						
ar						
L1.	0.8969204	1.081214	0.83	0.407	-1.22222	3.016061
L2.	-0.2895057	1.19927	-0.24	0.809	-2.640031	2.061019
L3.	0.6695635	0.6970543	0.96	0.337	-.6966378	2.035765
L4.	-0.1157677	1.058147	-0.11	0.913	-2.189698	1.958163
L5.	0.1784617	0.8733865	0.20	0.838	-1.533344	1.890268
L6.	0.0657371	0.7813321	0.08	0.933	-1.465646	1.59712
L7.	-0.4426692	0.650481	-0.68	0.496	-1.717589	.8322502
ma						
L1.	-488.4981	403891.6	-0.00	0.999	-792101.5	791124.5
L2.	639.8739	530186.2	0.00	0.999	-1038506	1039786
L3.	-271.7212	225393.7	-0.00	0.999	-442035.3	441491.9
/sigma	-196.9506	163284.5	-0.00	0.999	-320228.7	319834.8

Table 11

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	956.34

Table 12

ARIMA (2,1,2)

ARIMA regression

Sample: **2000m8 – 2008m12**
Log likelihood = **-755.5266**

Number of obs = **101**
Wald chi2 (**4**) = **133.68**
Prob > chi2 = **0.0000**

D2.TVP	Coef.	OPG Std. Err.	z	P> |z|	[95% Conf.	Interval]
TVP						
_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
/sigma	427.2118	15.14778	28.20	0.000	397.5227	456.9009

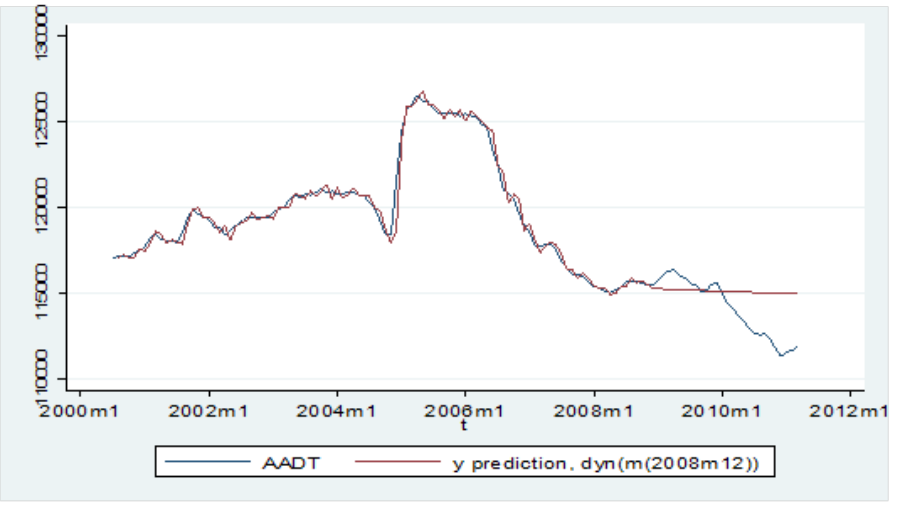


Fig. 4.

Actual and Predicted AADT for Lark Ellen

Table 13
Results Obtained from Different Methods for Target Year 1996

Method	Predicted value	Actual value	Error (%)
Trend Line Analysis	209702058	108336195	93.57
Econometric Analysis	115055054	108336195	6.202
Time Series Analysis	109951968	108336195	1.491

the three methods) is 1996. The comparative results of analyses done using different approaches are collated in Table 13.

Thus it can be seen that there was an unacceptably high error value for trend line analysis, resulting from an absurdly high overestimation. This invokes the need for a method that lends more dependability and is more logical to arrive at more acceptable results. For econometric analysis, the error value was significantly lower than that from Trend Line Analysis. This indicates that using a combination of a demographic and an econometric indicator, Population and Per Capita Income respectively, enables arriving at better approximations of future traffic demand than using a single indicator such as Gross National Product (GNP). As can be seen, the error level of result obtained from Time Series Analysis was considerably lower than that from the other two methods, reflecting its potential as a useful forecasting technique in the future. The results of Time Series Analysis on data from PeMS, DOT, California have partly been discussed along with illustrations in the same section where the analysis for each selected location has been dealt with (Section 4.2). The inferences drawn from these results have also been discussed in the same section. For Lark Ellen, the overall quality of results had been consistent and independent of the position of observation period with respect to the forecasting window. ARIMA (2,1,2) predicted an AADT value of 114959 for Mar. 2011 while the actual value is 111834 (resulting in an overestimation error of 2.794%).

6. Conclusion

Advancement in the methods of traffic forecasting is the need of the hour in India's context. As can be derived from the results obtained in this study, Trend Line Analysis usually results in high overestimation of future traffic volume. Improving on the accuracy of estimation is imperative to reasonable allocation of scarce resources like land, labor and funds, especially for developing nations. The study suggests that the use of more logical, dependable and advanced methods of analysis like Econometric Analysis and Time Series Analysis results in more acceptable results. Econometric modeling is able to better reflect the effect of two factors - growth in the total number of users as well as their purchasing power to use new developments in transportation facilities than the traditional Trend Line Analysis method which incorporates only the country's total productivity level (an indicator such as Gross National Product) for estimation of the number of prospective users. Time Series Analysis deserves a special mention. This method has been in use for short term forecasting in fields of finance and economics for a long time now and an understanding of its use in transportation engineering must be developed. The time frame for accurate forecasts by this method can be further investigated. As has been

seen, this method is very effective for short-term forecasts. For analysis done with IRC data using trend line and econometric approaches in addition to TS analysis, the error with time series modeling was considerably lower than the other two (1.49% for TS, 6.202% for econometric and 93.57% for trend line analysis) even though R^2 values were high for regression equations for the other two approaches. The level of its effectiveness over increasing time span can be checked for an acceptable degree of accuracy. If the limitation of high and rich data requirement for this method is overcome by implementation of proper technology over time, then in agreement with the findings of other researchers, it can be proposed to contribute favorably towards accurate traffic forecasting in time to come.

7. Limitations of the Study

The following are the limitations of the study:

- For India, the analyses have been done on the total vehicular population for the whole country. Due to data availability constraints for most of the variables used, analysis for a specific project level study remains to be performed.
- The study investigates the potential of Time Series Analysis majorly based on the data from California, US. Such extensive analysis on Indian data needs to be done to gauge its actual effectiveness in the Indian context.

8. Further Research

- For Trend Line Analysis, data spanning over a longer time frame should give an even better picture of its efficacy. If more data is made available for recent observation periods, then it will be interesting to observe the changes incurred in the results of analysis.
- For Econometric Analysis, various other combinations of demographic and/or economic indicators can be tried out to see the improvement in results with such changing combinations. Further, the level of change in accuracy with inclusion of three (or more) such variables in place of two can be researched and thus the marginal cost effectiveness for the inclusion of these variables can be checked.
- Availability of rich, comprehensive traffic data has been a major challenge in case of Time Series Analysis for India. The scope of study can be improved to a great extent if this limitation is overcome. Use of multivariate Time Series methods like GARCH (Generalized Auto Regressive Conditional Heteroskedasticity) and ARCH Processes, incorporating factors like change in land use patterns and a few relevant economic indicators may produce even more accurate results since this will provide the dual advantage of Time Series method and econometric modeling to a good extent. For this the data already cited from different sources may be of use.

Moreover, traffic forecasting is a field which works with a very dynamic variable. Therefore, many prospective methods, other than the ones taken up during this study can be an area of investigation. The scope and applicability of these methods should be kept in mind while doing such research. Emphasis must be laid on the usability of any such modeling technique for its application at specific project level work also. There may be a certain level of variation in results obtained in case of working with data for the whole country and the other one with data for specific project corridors since there may be not-so-apparent factors influencing the overall traffic demand in the latter case.

Acknowledgements

The authors would like to thank Dr. A. K. Giri, HoD, Economics & Finance Group, BITS, Pilani, and Mr. N. K. Jha, HoD, Economics, O.P. Jindal School, Patratu for their useful suggestions and contributions. They would also thank Jane Berner of Caltrans for providing access to the PeMS data and for her useful suggestions and illustrations about the database.

References

Baltagi, B.H. 1999. *Econometrics*, 2nd Edition. Springer. 105-116.

Bain, R.; Plantagie, J.W. 2004. *Traffic Forecasting Risk: Study Update 2004*, Infrastructure Finance, Standard & Poor's Ratings Direct. The McGraw-Hill Companies, New York. 99-132.

Bhar, L.M.; Sharma, V.K. 2005. *Time Series Analysis*. Indian Agricultural Statistics Research Institute, New Delhi: 1-15.

Box, G.E.P.; Jenkins, G.M. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day. 570-576.

Cervero, R. 2003. *Are Induced Traffic Studies Inducing Bad Investments?*, *ACCESS* 22: 22-27.

Cervero, R.; Hansen, M. 2002. Induced Travel Demand and Induced Road Investment: A Simultaneous Equation Analysis, *Journal of Transport Economics and Policy*, 36(3): 469-490.

Clark, S. 2003. Traffic Prediction using Multivariate Nonparametric Regression, Journal of Transportation Engineering, ASCE. DOI: http://dx.doi.org/10.1061/(ASCE)0733-947X(2003)129:2(161), 129(2): 161-168.

Dhingra, S.L.; Mujumdar, P.P.; Gajjar, R.H. 1993. Application of time series techniques for forecasting truck traffic attracted by the Bombay metropolitan region, *Journal of Advanced Transportation*. DOI: http://dx.doi.org/10.1002/atr.5670270303, 27(3): 227-249.

Economic Survey, 2007-08. 2008. Govt. of India. Published by Oxford University Press, New Delhi.

Gujarati, D.N. 2004. *Basic Econometrics*, 4th Edition. The McGraw-Hill Companies, New York. 448-547.

Green, W.H. 1993. *Econometric Analysis*, 2nd Edition. Macmillan Publishing Company, New York. 535-538.

Hansen, B. 2007. Time Series Analysis on STATA. Available from Internet: <http://www.ssc.wisc.edu/~bhansen/390/stata.pdf>.

Hymel, K.M.; Smalla, K.A.; Van Dender, K. 2010. Induced Demand and Rebound Effects in Road Transport, *Transportation Research Part B: Methodological*. DOI: http://dx.doi.org/10.1016/j.trb.2010.02.007, 44(10): 1220-1241.

India: Economic Indicators (1951-2001). Available from Internet: <http://www.indiastat.com>.

Kadiyali, L.R. 1987. Road Transport Demand Forecast for 2000 AD, *Journal of the Indian Roads Congress,* Paper no. 384, 48(3).

Kadiyali, L.R.; Shashikala, T.V. 2009. Road Transport Demand Forecast for 2000 AD Revisited and Demand Forecast for 2021, *Journal of the Indian Roads Congress*, Paper no. 557, 70(3): 235-237.

IRC: SP-30. 1984. Manual on "Economic Evaluation of Highway Projects in India, *Indian Roads Congress*, New Delhi.

Matas, A.; Raymond, J.L.; González-Savignat, M.; Ruiz, A. 2009. Demand Forecasting in the Evaluation of Projects. Working Paper in *Economic Evaluation of Transportation Projects*. 1-31.

Moeinaddini, M.; Asadi-Shekari, Z.; Shah, M.Z. 2012. The Effectiveness of Private Motorized Trips Indicators in Reducing Car Usage, *International Journal for Traffic and Transport Engineering*. DOI: http://dx.doi.org/10.7708/ijtte.2013.3(2).05, 2(4): 347-358.

Multiple Linear Regression. Available from Internet: <http://www.ccsr.ac.uk/publications/teaching/mlr.pdf>.

Nanda, P.K. 2005. Road Project Appraisal Process in India, Country Project Report. *Central Road Research Institute*, CSIR, New Delhi. 17-37.

Naylor, T.H.; Seaks, T.G.; Wichevn, D.W. 1972. Box-Jenkins methods: an alternative to economic forecasting, *International Statistical Review*, 40(2): 123-137.

Nelson, C.R. 1973. *Applied Time Series Analysis: For Managerial Forecasting.* San Francisco: Holden-Day. 139-169.

Nihan, N.L.; Holmesland, K.O. 1980. Use of the Box and Jenkins Time Series Technique in Traffic Forecasting, *Transportation*, 9(2): 125-143.

Oswald, R.K.; Scherer, W.T.; Smith, B.L. 2001. *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.

Pankratz, A. 1983. *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. John Wiley & Sons, New York. 234-370.

Parthasarathi, P.K. 2001. *A Disaggregate Analysis of Induced Demand*, M.S. Thesis. University of Minnesota, Twin Cities.

Project: Feasibility for 6-laning of NH-2 from Delhi-Agra Project on DBFO pattern under NHDP Phase V (October 2007), *Consulting Engineering Services*, Chapter 3.

Ramsey, S. 2005. Of Mice and Elephants, *ITE Journal*, 75(9): 38-41.

Skamris, M.K.; Flyvbjerg, B. 1997. Inaccuracy of Traffic Forecasts and Cost Estimates on Large Transport Projects, *Transport Policy*. DOI: http://dx.doi.org/10.1016/S0967-070X(97)00007-3, 4(3): 141-146.

Time Series Data on Road Transport Passenger and Freight Movement (1951-1991). 1996. Special Publication 45, *Indian Roads Congress*, New Delhi.

MODELIRANJE TRENDA RASTA I TEHNIKA PROGNOZIRANJA POPULACIJE VOZAČA U INDIJI

Kartikeya Jha, Nishita Sinha, Shriniwas Shrikant Arkatkar, Ashoke Kumar Sarkar

Sažetak: Prognoziranje i procena porasta populacije vozača predstavlja neophodnu meru svakog značajnijeg razvoja saobraćajnog inženjerstva, koja zahteva uočavanje trenda u prošlosti i njegovo korišćenje u cilju predikcije budućeg trenda na osnovu odgovarajućih pretpostavki, simulacija i modela kreiranih pomoću promenljivih. Ovaj rad ima za cilj da razmotri pristupe korišćene u praksi i da ispita savremeniji pristup – analizu vremenskih serija. U radu su prikazane tri fundamentalno različite metode i upoređeni su dobijeni rezultati kako bi se proverio odgovarajući nivo tačnosti u prognozi populacije vozača za istu ciljnu godinu. U okviru sprovedene studije, rezultati dobijeni pomoću analize vremenskih serija su se pokazali znatno preciznijim u odnosu na one dobijene pomoću analize linije trenda i primetno boljim od rezultata dobijenih pomoću ekonometrijske analize. U cilju verifikacije dobijenih rezultata, urađen je drugi skup analiza sa novijim ulaznim podacima preuzetim od AADT podataka iz PeMS, Kalifornija. Između ostalog, verifikacija je sprovedena bez dodatnih statističkih poboljšanja koja se dobijaju kada se analiza vremenskih serija primenjuje sa potpunim i tačnim podacima. Na osnovu svih skupova korišćenih podataka i lokacija koje su analizirane pri prognoziranju, u radu je utvrđeno da analiza vremenskih serija predstavlja moćan alat za prognozu.

Ključne reči: ekonometrija, vremenske serije, ARIMA, regresija, elastičnost, srednja kvadratna greška.