DOI 10.20544/HORIZONS.B.05.1.18.P09
UDC 656.05:303.714(497.11)"2004/2019"

DEVELOPMENT OF TRAFFIC VOLUME FORECASTING MODEL USING MULTIPLE REGRESSION AND PRINICPAL COMPONENT ANALYSIS

Ramadan Duraku
University of Prishtina"Hasan Prishtina"-FME-Department of
Traffic and Transport, Prishtina, Kosovo
ramadan.duraku@uni-pr.edu
Vaska Atanasova
University "St.Kliment Ohridski"-Faculty of Technical Sciences
Department of Traffic and Transport, Bitola, Macedonia
vaska.atanasova@tfb.uklo-edu.mk

ABSTRACT

This paper is focused in development model for traffic volume forecasting in Anamorava region. Demographic and socioeconomic macro variables (independent variables) are identified at country and region level which have an impact on generation of traffic volume (dependent variable), using dataset for the period 2004-2016. Multiple Regression Analysis (MLR) method was used to build dependency between variables and model development. In order to increase forecasting capabilities of the MLR model, it was necessary to eliminate high correlation between variables (multicollinearity phenomenon). A new methodology was used, with included Principal Component Analysis (PCA) by transforming original variables in Principal Components (PCs). With the combination PCA and MLR methods are developed hybrid model as called Principal Component Regression (PCR). By employing performance indicators, it was found that the PCR model performs much better than MLR model.

KEY WORDS: traffic volume; forecasting model; methods; MLR; PCR.

INTRODUCTION

Traffic volume generations comes as a result of increased demand of population for conducting trips for carrying out their needs and activities. Since this dependence exists, then the purope of this paper is to develop a model for traffic volume forecasting in Anamorava region with high capability or less error in forecasting [1]. Based on review literature [2],[3] there was clear idea on which variables are required for completion of analysis. By identifying demographic and socioeconomic variables at country and region level-independent variables $X=(X_{i1},X_{12},...X_{ki})$ and the traffic volume expressed as annual average daily traffic (AADT) at regional level–dependent variable (Y_i) and after build the dataset for the period 2004-2016 received from various state competent institutions [4],[5] and [6] it was possible to start to develop the model.

METHODOLOGY

MLR method was initially used to develop the model. As a result of high correlation between indipendent variables appeared multi-colinearity phenomenon, which was an drawback to the inclusion of all original variables in the forecasting model [7]. To eliminate this weakness and develop the model with best performances, PCA method was used by transforming the original independent variables into PCs. Combined PCA and MLR methods provide hybrid PCA-MLR model or PCR model [8].

A. STUDY AREA

The study is conducted on the main road network in Anamorava region within Kosovo, more specifically in four locations in which Automatic Traffic Counting (ATC) were installed for registering data for traffic volume: 1-Slivovo, 2-Sojevo, 3-Ranillug and 4-Pasjan as presented in fig. 1.

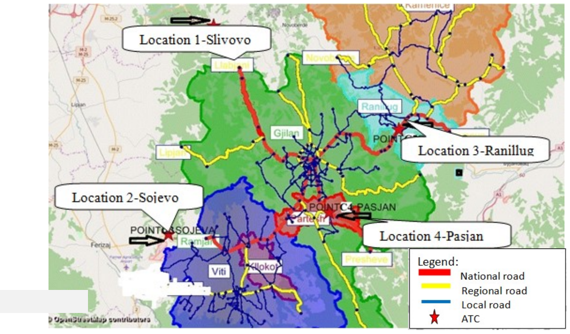

Fig.1. Study area and locations of ATC in Anamorava region

Testing homogeniety of data for traffic volume for four locations are done using Levene test=0.827 with the significance level (p=0.429>0.05). This proves that the variability of traffic volumes for four locations has approximate values. Based on this reason, in this paper is treated only Slivovo location which also represents other locations.

B.METHODS USED
B1.Principal Component Analysis

PCA is a multivariate statistical nonparametric method which is used for summarization of collected information by some observed variables which are correlated and by transforming them to a number of non-correlated PCs [9]. This approach enable the elemination o multi co-linearity [10]. The PCs gained by the combination of original variables, in ordering from the largest to the smallest. Values of each PCs can be calculated by using Eq.1:

$$PC_i = a_{i1} \times X_1 + a_{i2} \times X_2 + ... + a_{in} \times X_n = \sum_{j=1}^{n} a_{ij} \times X_j \qquad \textbf{(1)}$$

In which $X_i = X_1, X_2, ..., X_n$ are n original variables in the data set; a_{ij} are eigenvectors, where $i=1,2...n; j=1,2...n.$

In order that PCA to be effective, three basic steps are suggests: estimation of the suitabilty of data, extracting the main components, rotation of vectors. Suitability of the data is carried out using the Kaiser-Meyer-Olkin (KMO) test with the significance according to Bartlett's test of sphericity [11]. This test can be calculated by Eq.2:

$$KMO = \frac{\sum_{i'j} \sum r_{ij}^2}{\sum_{i'j} \sum r_{ij}^2 + \sum_{i'j} a_{ij}^2} \qquad \textbf{(2)}$$

In which r_{ij} is a simple correlation, while a_{ij} is partial correlation between i and j. Since it is verfied that (KMO>0.5) with significance (p<0.05) is fullfilled then the PCA justified. The procedure continues by extracting PCs and calculating the eigenvalues from the covariance matrix. PCs that have eigenvalues more than 1 (eigenvalues>1) should be considered when developing the model [12]. Finally, the rotation of factors is used, where new factors are gained and can be named and also interpreted.

B2. Multiple Regression Analysis

MLR is one of statistical methods which is used to find out relation between one dependent variable (Y_i) and some other independent variables (X_{ni}). The general form of MLR model is given by Eq. 3 [13]:

$$Y_i=\beta_0+\beta_1\times X_{1i}+...+\beta_k\times X_{ni}+\varepsilon_i \qquad \textbf{(3)}$$

where β_i regression coefficients, X_{ni} dependent variables and ε_i error term which is supposed to be normal distribution around zero and constant variance. It is also supposed the residuals should not correlate. For calculate the values of parameters the Least Square Method (LSM) is used.

B3. Principal Component Regression

Analysis through PCA-MLR provide a combination of PCA and MLR methods to form a reciprocal relationship between the dependent variable (Y_i) and the *PCs* which are obtained as a result of the multiplication of independent original variables with eigenvectors [14]. This combination enables the creation of a hybrid method known as PCR. Here PCs obtained by PCA method used as independent variable in MLR model. The general form of PCR model is given by Eq. 4:

$$Y_i=\beta_0+(\beta_1\times PC_1+\beta_2\times PC_2+.....+\beta_n\times PC_n)=\beta_0+\sum_{j=1}^{n}\beta_j\times PC_n+u_i \qquad \textbf{(4)}$$

where β_0 intercept; $\beta_1, \beta_2..., \beta_n$ regression coefficients; $PC_1, PC_2..., PC_n$ principal components; and u_i is error associated during regression.

RESULTS AND DISCUSION

The analysis begins by descriptive statistics for variables, table 1. The values for all variables are given, with exception of X_1 variable which indicates years and all other variables are related according to years.

Table 1. Descriptive statistic for variables reviewed

Variables			N	Min	Max	Mean	Std. dev
	Traffic volume	Y	1	6325	10439	7449	1240
State level	Population	X_2	12	1786282	1891906	1846303	37473
	Household	X_3	12	278915	338618	311062	17863
	Employment	X_4	12	236181	340911	290450	33631
	Vehicleregistration	X_5	12	179157	336942	249102	52192
	CPI	X_6	12	77	101	90	9
	GDP	X_7	12	3006100	5984900	4431790	1072114

	Per capita income	X_8	12	1763	3356	2507	562
	Gasoline price	X_9	12	0.840	1.160	1.015	0.027
Region	Population	X_{10}	12	240502	254723	248583	5045
	Household	X_{11}	12	48999	51442	50504	685
	Employment	X_{12}	12	32270	43692	37302	3983
	Vehicleregistration	X_{13}	12	29031	53806	39419	8514

Estimation and model development are done using two methods in above mentioned and employing SPSS software.

RESULTS BY MLR

By MLR method, the relation between independent variables and traffic volume is found out. The MLR model is developed through 13 observations (years) and by 12 independent variables (X_{ni}) against dependent variable (Y_i). Using stepwise technique, the best model is found.

Table 2. Model summary of statistical parameters

Model Summary[a]					
Model	R	R^2	Adjusted R^2	Std. Error	D.Watson
1	0.933[a]	0.871	0.859	465.53011	1.338
ANOVA					
Model	SumSquares	Df.	Mean Square	F	Sig.(p<0.05)
Regression	16071147.6	1	16071147.643	74.157	0.000[b]
Residual	2383901.12	11	216718.284		
Total	18455048.76	12			
Coefficients`					
	B	Std.Err	t	Tolerance	VIF	Sig.(p<0.05)
Constant	2091.432	635.437	3.291			0.007
X_{13}	0.136	0.016	8.611	1.000	1.00	0.000
a. Dependent Variable: Traffic volume (Y)					

Based on table 2, value of determination coefficient R=0.933, respectively 93.3% confirmed that dependent variable has high correlation with other independent variables. The value of $R^2=0.871$ respectively 87.1% means that the dependent variable is explained by independent variables. In addition, value of adjusted $R^2=0.859$ show that 85.9% of variance of dependent variable is explained by independent variables. According to VIF=1<10, is verified that the model does not suffer from multi-co linearity. Also, value of Durbin-Watson (DW=1.338) shows does not face problem with first rule of auto correlation. The model resulted statistically significant by F-test (F=74.157) and t-test (t=8.611) for the level of (p<0.05). Also, indicates that residuals are distributed in normal way by zero mean and constant variance. The model by MLR method is given by Eq.5:

$$Y=2091{,}432+0{,}136\times X_{13} \quad \textbf{(5)}$$

The regression coefficient before X_{13} variable is positive which means that by increasing the level of motorization in level of Anamorava region, the value of dependent variable “traffic volume” is also increasing.

RESULTS BY PCR

Development model for forecasting traffic volumes here is by combine MLR and PCA methods using 12 different original variables, which will be combined with some other general factors afterwards. Based on values of KMO=0.720>0.5 and Bartlett's Test ($p<0.05$), confirmed that the dataset support factorial analysis and suggests that independent variables may be grouped into smaller group of impact factors called PCs. These PCs then could be using as input variables in MLR method.

Table 3. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.720
Bartlett's Test of Sphericity	Approx. Chi-Square	292.764
	df.	55
	Sig.	0.000

Table 4 contains eigenvalues associated to each component linear (factor) before extraction, after extraction and in the phase of rotation. Before the phase of extraction there are 12 linear components identified within the data set. By using PCA method all variables are extracted in two factors with eigenvalues bigger than 5% which are named PC1 and PC2. Further, other factors contribute much lower in eigenvalues and are close to zero which means there is presence of multi co-linearity. Since the aim of application of PCA is to eliminate the problem of multi co-linearity, other factors PCs have been ignored in further analyses, where only PC1 and PC2 remain.

Table 4. Total variance explained by each PCs

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of SSL
	Total	% of Variance	Cumu %	Total	% of Variance	Cum %	Total
1	**10.865**	**90.538**	**90.538**	**10.86**	**90.538**	**90.53**	**10.285**
2	**0.678**	**5.647**	**96.185**	**0.67**	**5.647**	**96.18**	**9.234**
3	0.191	1.592	97.777				
4	0.101	0.839	98.616				
5	0.082	0.685	99.301				
6	0.058	0.482	99.783				
7	0.019	0.158	99.941				

8	0.004	0.035	99.977				
9	0.002	0.014	99.990				
10	0.001	0.007	99.998				
11	0.000	0.002	100.000				
12	3.305E-11	2.754E-10	100.000				

The eigenvalues associated to these factors PC1 and PC2 are shown in the table 4 of extraction like Sums of Squared Loading (SSL) with the variability about 96%. In the last column of table 4 are presented factors after application of rotation process. Through the rotation process it is possible to optimize structure of the factors, where the relative weighting is also given. Thus, the first factor takes approximate value (from 10.865 to 10.285), while the value for the second factor varies drastically (from 0.678 to 9.234). Correlation between original variables and new PCs are presented by Table 5. Only, high values (value in bold) corresponding to respective PCs taken into account. Results from table 5, show that all variables have higher values (higher than 0.9) for each PCs. Thus, variables X_{12}, X_5, X_7,X_{13}, X_4, and X_6 have high impact in PC1, while variables X_2,X_{10},X_9,X_{11},X_3 have high impact in PC2. The high value of load was used to get coefficients for each PC in forming points of coefficients to PC score coefficient, Table 6.

Table 5. Component loading structure matrix

Original variables	Component	
	PC1	PC2
X_{12}	**0.988**	0.825
X_5	**0.988**	0.779
X_8	**0.986**	0.845
X_7	**0.985**	0.865
X_{13}	**0.984**	0.721
X_4	**0.977**	0.797
X_6	**0.939**	0.925
X_2	0.907	**0.957**
X_{10}	0.907	**0.957**
X_9	0.656	**0.948**
X_{11}	0.834	**0.943**
X_3	0.902	**0.925**

Table 6. Component score coefficient matrix

Original variables	Component	
	PC1	PC2
X_2	0.030	0.177
X_3	0.044	0.150
X_4	0.156	-0.037
X_5	0.172	-0.064
X_6	0.066	0.120
X_7	0.125	0.023
X_8	0.136	0.002
X_9	-0.113	0.379
X_{10}	0.030	0.177
X_{11}	-0.005	0.224
X_{12}	0.147	-0.019
X_{13}	0.200	-0.118

Multiplying scores of coefficients (eigenvectors) and values of original variables obtained scores for each PC. These scores are used as independent variables in MLR analysis to determine all significant PCs which could be used in the model. As noted, and reflected in table 6, all variables are covered by two selected PCs supported by the equations Eq.6 and Eq.7:

$$\begin{aligned}PC1&=0{,}030\times X_2+0{,}044\times X_3+0{,}156\times X_4+0{,}172\times X_5+0{,}066\times X_6+0{,}125\times X_7\\&+0{,}136\times X_8-0{,}113\times X_9+0{,}030\times X_{10}-0{,}05\times X_{11}+0{,}147\times X_{12}+0{,}200\times X_{13}\end{aligned} \quad \mathbf{(6)}$$

$$\begin{aligned}PC2&=0{,}177\times X_2+0{,}150\times X_3-0{,}037\times X_4-0{,}064\times X_5+0{,}120\times X_6+0{,}023\times X_7\\&+0{,}02\times X_8+0{,}379\times X_9+0{,}177\times X_{10}+0{,}224\times X_{11}-0{,}019\times X_{12}-0{,}118\times X_{13}\end{aligned} \quad \mathbf{(7)}$$

The PCR model is developed through 13 observations (years) by two independent variables (PCs) against dependent variable (Y_i). Using stepwise technique, the best model is found.

Table 7. Model summary of statistical parameters using PCR

Model Summary[a]						
Model	R	R^2	Adjusted R^2	Std. Error	D.Watson	
1	0.948[a]	0.899	0.879	431.58135	1.486	
ANOVA						
Model	SumSquares	df	Mean Square	F	Sig.(p<0.05)	
Regression	16592424.14	2	8296212.07	44.540	0.000[b]	
Residual	1862624.62	10	186262.46			
Total	18455048.76	12				
Coefficients[a]						
	B	Std.Err	t	Tolerance	VIF	Sig.(p<0.05)
Constant	7449.308	119.699	62.234			0.000
PC1	1516.286	205.210	7.389	0.369	2.713	0.000
PC2	-473.241	205.210	-2.306	0.369	2.713	0.044
a. Dependent Variable: Traffic volume (Y)						

The results obtained in table 7, show that the dependent variable has high correlation with the acquired PCs that are treated as independent variables at 94.8%, respectively R = 0.948. While $R^2 = 0.899$ indicates that 89.9% of the dependent variables are explained by uncorrelated PC1 and PC2.

In addition, value of adjusted $R^2=0.879$ show that 87.9% of variance of dependent variable is explained by PC1 and PC2. Model which is developed by PCR method, resulted significant according by F-test (F= 44.540) and t-test (t=7.839 and t=-2.306) with the level of ($p< 0.05$).

Results in table 7, also show clearly there is no problem with multi co-linearity because the value of VIF is much smaller than 10 (VIF=2.713<10) and the value of Durbin-Watson (DW=1.486) does not face problem with first rule of auto correlation.

Also, indicates that residuals are distributed in normal way by zero mean and constant variance. The model by PCR method is given by Eq.8:

$$Y=7449{,}308+1516{,}286\times PC1-473{,}241\times PC2 \quad \mathbf{(8)}$$

From Eq. (8) it is seen that the PC1 component has a positive impact while the PC2 component has a negative impact on traffic volumes.

In addition, the results obtained by the MLR and PCR methods are presented graphically as in Figure 2, where the black color present the traffic volume measured by ATC, whereas forecasting by the methods (MLR-blue, PCR-red).

By comparing the results, it is seen that the PCR method give close forecats, because the red curve has more approximate values with the black curve.

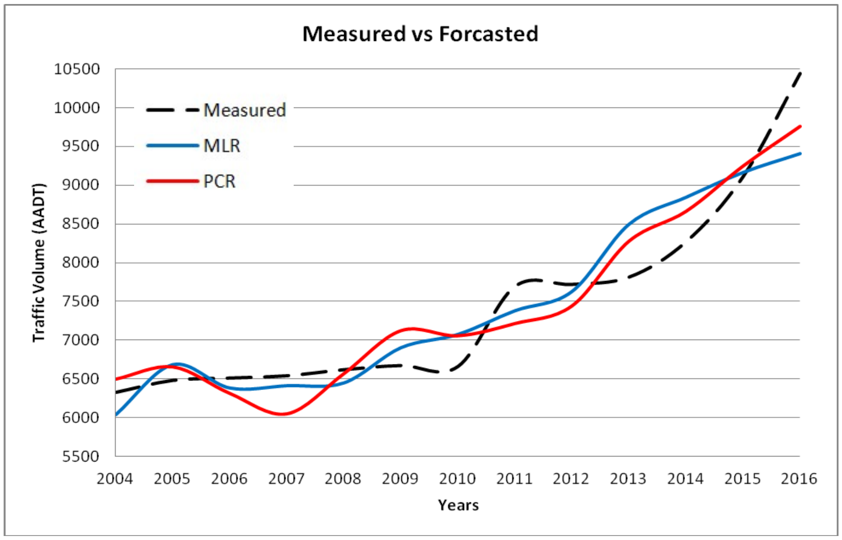

Fig. 2. Comparison of measured vs. forecated traffic volume

PERFORMANCE INDICATORS

In order to have a better overview which of the methods gives the most accurate results in traffic volume forecasting, the comparison was done by indicators' performance [15].

In table 8, in summary way are given values of these indicators. All the indicators show that the model which was developed according to the PCR method guarantee in better forecasting than MLR.

Table 8. Performance indicators between MLR and PCR models

Performance Indicators	MLR	PCR
Adjusted R^2	0.859	0.879
Mean Absolute Deviation (MAE)	361	336
Root Mean Squared Error (RMSE)	428	370
Root Mean Square Percentage Error (RMSPE)	5.09	4.93

CONCLUSION

In this paper is given a development model for traffic volume forecasting according to MLR and PCR methods for the Anamorava region. For the development of this model, data set in time series for 12 macro variables (independent) and traffic volume(dependent variable) were taken into account. With the use of these variables and regression-based methods, several significant models were drawn.

It was proven that MLR method has weakness in that as a consequence of high correlation, as a result number of independent variables drastically falls in model development. While, the PCR method has the advantage in the elimination of multicollinearity, but as a weakness is in interpretation between PCs and original variables. By comparing graphical results and performance indicators, it was found that PCR model is better than MLR model because it gives smaller error in traffic volume forecasting.

Finally, we conclude that this model can be used for traffic volume forecasting in case of planning transport strategy in this region.

REFERENCE

1. EgisBceom&COWI, National model of transport for Kosovo, Prishtina, February 2009.
2. Ortuzar J.D, Williamson L.G, Modelling Transport, Third Edition.
3. Fricker J.D, Saha S.K, Traffic volume forecasting methods for rural State Highways, Final Report. Joint highway research project FHWA/IN/JHRP-86/20, 1987.
4. Slipunas T., Annual average daily traffic forecasting using different techniques, Transport and Road Research Institute. Vol. XXI, No.1,38-43, Published by Transport, Lithuania, 2006.
5. Agency Statistics of Kosovo, General Statistics of Kosovo, Prishtina, 2016.
6. Regional Development strategy for the economic region east, Implemented by RDA. Prishtina, 2010.
7. Duraku R, Vaska A., Evaluation for long term traffic volume forecasting using multiple regression analysis and principal component regression models, Transport for today's society, 2nd

International Conference, Bitola, Republic of Macedonia, 17-19 May 2018.

8. Slipunas T., Annual average daily traffic forecasting using different techniques, Transport and Road Research Institute. Vol. XXI, No.1,38-43, Published by Transport, Lithuania, 2006,
9. Keho Y., The basics of linear principal components analysis. Ecole Nationale Superieure de Statistique et d'Economic Appliquee (ENSEA), Abidjan Cote d'Ivoire.
10. Paul R.K., Multi co linearity: Causes, effects and remedies. I.A.S.R.I, Library Avenue, New Delhi.
11. Kalayaci S., SPSS in Practice. Statistical Techniques with Many Variables, Six Editions, 2014.
12. Sousa et al., Multiple linear regression and ANN based on principal components to predict ozone concentrations. Elsevier, 2006.
13. Washington S.P, Karlaftis M.G, Mannering F.L., Statistical and Econometric methods for transportation data. Second Edition, Taylor and Francis Group, LLC, 2011.
14. P.B Mistry., Principal regression for crop yield estimation. Springer, 2016.
15. Saha K.Sunil and Fricker J.D.,The development of a procedure to forecast traffic volumes on urban segments of the state and interstate highway systems. Final report FHWA/IN/HJHP-90/11, Joint Highway Research Project. School of Civil Engineering, Purdue University, 1990.