

TRAFFIC PRODUCTION MODEL USING MULTIPLE REGRESSION ANALYSIS AND RADIAL BASIS FUNCTION NEURAL NETWORK

*Prof.dr.Vaska Atanasova.¹, Mr.sc. Ramadan Duraku.², Associate prof.dr Beti Angjelevska³, Prof.dr.Nikola Krstanoski.⁴
 Faculty of Technical Sciences, University "St.Kliment Ohridski", Bitola, Macedonia^{1,3,4}
 Faculty of Mechanical Engineering, University of Prishtina "Hasan Prishtina", Prishtina, Kosovo²
 (* Corresponding author: vaska.atanasova@tfb.uklo-edu.mk)

Abstract: In order to establish this model, several independent variables of aggregate level and dependent variables have been taken into account for traffic production of a region for a period of 24 hours. The objective of this paper is to identify and estimate the main variables which are significant for development of a suitable model for forecasting the traffic production for Anamorava region. Establishment and assessment of a model will be done using techniques such as Multiple Regression Analysis (MRA) and Radial Basis Function Neural Network (RBFNN). The results gained according to these two techniques will be compared to find out differences. In this regard performance analysis through indicators R^2 , ME, RSME which are also used to show errors in forecasting. The results are in favour of RBFNN model because it performs better than MRA in forecasting traffic production.

Keywords: TRAFFIC PRODUCTION, MODEL, MRA, RBFNN

1. Introduction

Transport planning process which is closely related to travel demand forecasting is essential for the design of transportation facilities and services, and also for planning, investment and policy making [1]. The models of forecasting in transport planning are very important to foresee increase of traffic and trips in the future. [2]. The traffic production modelling is to be done by finding relation between traffic countings which exit from a zone from origin - destination survey result and other independent variables of macro level in a given zone. This was a reason to undertake field research to find out the matrix origin-destination of traffic production for a period of 24 h treated here like independent variable in the Anamorava region. Later by application of various modelling techniques like those traditional –linear regression and those non traditional-RBFNN it will be possible to built a model for forecasting of traffic production.

1.1 The Objectives of the study

The objectives of this paper are: 1. Development of a suitable model for forecasting traffic production in Anamorava region using techniques such as MRA and RBFNN, and (2) comparison of results according models MAR and RBFNN by application of performance indicators like R^2 , ME and RSME.

2. Materials and methods

The methodology of traffic forecasting in this paper is addressed in three parts: (a) Data collection, (b) Regression models and (c) ANN models.

3.1 Data collection

In order to establish the model, the requirement was to identify and collect data on independent variables as well as dependent variables. Collection of data for variables which will be incorporated in setting up models is done according to two ways explained as follows.

2.1.1 Aggregate data

Demographic, economic and terrain factors have been taken like independent variables in aggregate level of more impact. Relevant data on this issue have been taken from competent authorities for collection, maintenance and archiving them [3], [4], [5] etc. Selection of independent variables is done in line with relevant literature [6], [7], [8] and [9]. Exclusion from this is on some territorial variables such is distance and time and the other one which is the result of them namely, cost of trip which is done by an author based on primary information available in Google maps etc.

2.1.2 Field data for traffic counting

The key place in this paper is the traffic produced on daily basis by this region qualified like independent variables (Y). Data on them is found out through the research conducted specifically for this purpose. The outcome of this is a matrix O-D of trips which is very important to proceeds with further steps towards establishment of a relation between variables and finally to establish models. This is made possible by completing manual counting and interviews in cordon in eight selected locations in the main road network with higher traffic flow in region. Forms and special questionnaires are prepared in advance to complete counting and interviews, which contain direction of trip (O-D), purpose of trip, frequency of trips, number of passengers in vehicle etc. Depending on origin and the purpose of trip we took into consideration internal-external, external-internal, internal-internal and transit traffic. The traffic within one zone was not subject of research. Counting and interviews are completed simultaneously on 18.05.2016 and 21.05.2016 non-stop for 12 hours starting from 07:00 am until 19:00 pm. There are 11523 interviews completed which is 19.43% of total traffic flow in two directions which also show 59317 vehicles counted manually. Counting and interviews are conducted assisted by students of Transport and Traffic Department supported by Kosovo Police.

2.1.3 Data Processing

After processing of data it was possible to establish the matrix O-D of real trips for each category of vehicles. Thus, knowing number of traffic flows and interviews conducted for a period of 12 hours and for only two days of the week, which is one working day and the other day is weekend, it was not possible to further continue without conducting additional calculation coefficients in relation to K_k , K_{int} and $K_{convert(12h/24h)}$ in order to gain final matrix of daily trips by vehicles. After counting is completed there is a need to correct them. Counting of traffic shows we have various types of vehicles which influence the traffic and requirement is to convert vehicles per unit in order to have uniform traffic influence. This measurement based on categories per vehicle will be converted to vehicles per unit equal to a car based on the equivalent coefficient value $K_k=1.4$. Gained values need to be multiplied with this coefficient to make the equivalent of PCSE.

2.1.3.1 The coefficient of interview

It is obvious that technically is not possible to interview every vehicle driver. In order to find the number of trips by vehicle within the period of 12 hours coefficient of interview K_{int} was used. This is gained comparing the traffic flow and the interviews conducted within given period of time for each location separately as given hereunder in equation 1 and results are presented in Table 1.

$$K_{int} = \text{No.vehicles}_{12h} / \text{No.interviews}_{12h} \quad (1)$$

Table 1: Interview coefficient

Location	Traffic volume veh/12h	Number of interview	%	Interview coefficient of vehicles (K_{int})
Slivovo	5494	1131	18.01	4.86
Sojevo	5981	1092	15.42	5.48
Begrac	3137	880	24.23	3.56
Kllokot	4749	925	17.73	5.13
Dheu i Bardhe	1148	278	57.43	4.14
Ranillug	5350	894	16.47	5.99
Livoc i Poshtem	7173	1354	18.41	5.30
Muqibabe	571	115	59.62	4.96

By using this coefficient, it was possible to establish of O-D trip matrix for total number of vehicles for a time interval of 12 hours applying equation 2 as follows:

$$\text{Matrix}_{O-D} = \text{No.interviews}_{12h} \cdot K_{int}[\text{veh}/12h] \quad (2)$$

2.1.3.2 Converting traffic from 12 h to 24 h

After processing the traffic flow conducted from automatic mobile counters for an interval of 24 hours which have been installed in four locations through it was possible to get converting coefficient of traffic from 12 hours to 24 hours through equation 3.

$$K_{\text{convert}(12/24)} = \text{No.vehicles}_{12h} / \text{No.vehicles}_{24h} \quad (3)$$

Comparison of data for four locations resulted in its average value is $K_{\text{convert}(12h/24h)} = 1.5$. In this regard, application of traffic flow values and values gained of these coefficients in equation 4, the final matrix is gained on O-D of trips of vehicles for the period of 24h.

$$\text{Matrix}_{O-D} = \text{Volume}_{12h} \cdot K_k \cdot K_{int} \cdot K_{\text{convert}12/24}[\text{veh}/24h] \quad (4)$$

This results from matrix is like an average of two days when counting and interviews are conducted with an intention not to require application of weekly nonlinear coefficient of trips.

4.2 Linear regression

In order to develop traffic production modelling, respectively to investigate relationship between of dependent variables Y_i and independent variables X_i , MRA technique is applied and this general form according to [10] is given in equation 5:

$$Y = \beta_0 + \sum_{i=1}^n \beta_i \cdot X_i + \epsilon_i \quad (5)$$

Where: Y_i - is number of productive trips per day, X_i -explanatory variable from 1 to n, β_0 - is regression constant; β_i -is regression coefficient and ϵ_i - error.

Correlation coefficient is verified between every dependent variable with each independent variable in a specific way. According to this it results that demographic, economic and territorial phenomenon of various zones are well correlated ($R > 0.5$) with total number of trips. In this way we decided that each of them to include in analysis. The model is done using SPSS package software. Stepwise technique, backward and automatic linear regression were used to get the most statistical significance equations, between independent variables on dependent variable in the model [11]. After several attempts it was not possible to get the best results according to only "lin-lin" model and we needed to intervene in transformation of data in dependent variable applying several permitted ways of transformation shown in wide literature on statistics and econometric functions such as "log" [12]. Based on evidence it resulted that application of "log" in transformation of dependent variable provide best results compared to other functions. Selection of variables in final equation of regression is done following some principles as mentioned following: a. non existence of multi co-linearity between variables, b. variables need to be in harmony with engineering and logic judgement in relation to

developments, and c. selection of models with the less possible number of variables, to have in possession a history of series of variables taken for review, to have high value for determination of coefficient R^2 and low value of mean error (ME), root mean square error (RSME) [13].

5.3 ANN models

Artificial Intelligence technique is more frequently used in the last years for forecasting in many areas. Part of this technique is fuzzy logic, neural network, genetic algorithms, etc. [14]. All these three techniques are used for solving complicated problems at the same time and in combination. There is also a tendency to use ANN which is a tool to solve problems, which makes possible processing of data like input-output and the results are guaranteed in treating some inaccuracies while analytic equation of input-output are not needed.

ANN is a calculation model which is defined through the type of neuron, connection architecture, algorithm for learning etc. Based on connection architecture there are [15]: a. Feed forward which consist of: MLP, PNN, GRNN, CFN, RBF, AMN and b. Feedback which consist of Jordan Elman network-JENN, Hopfield network-HN, Adaptive resonance theory-ART, Recurrent network-RNN etc. Each of them has its own specifics and similarities to each other.

In this paper, the radial basis function neural network (RBFNN) model is provided. The methodology and parts of RBFNN models, variables and analysis used are described in the following discussion.

2.3.1 The topology of RBFNN

A radial basis function (RBFNN) is a feed-forward, supervised learning network with only one hidden layer, called the radial basis function layer. The RBF network is a function of one or more predictors (also called inputs or independent variables) that minimizes the prediction error of one or more target variables (also called outputs). Predictors and targets can be a mix of categorical and scale variables. The topology of RBFNN consists of three layers: input layer, hidden layer and output layer (see Fig.1). This RBFNN proposed here is a special case of multilayer feed forward neural networks, but different in terms of node characteristics and learning algorithms [16].

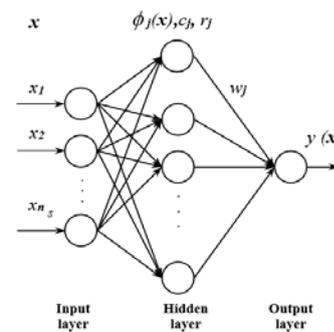


Fig 1. Architecture of radial basis function neural networks.

There is no calculation in input layer nodes. The input layer nodes only pass the input data to the hidden layer. The input layer consist of n_i nodes where input vector $x = (x_1, x_2, \dots, x_{n_i})$. The hidden layer consists of n nodes and each hidden node $j = 1, 2, \dots, n$ has a center value c_j . Each hidden layer node performs a nonlinear transformation of the input data onto new space through the radial basis function. The most common choice for the radial basis function is a Gaussian function, given by equation 6:

$$\Phi_j(x) = \exp(-\|x - c_j\|^2 / r_j^2) \quad (6)$$

where: $\|x - c_j\|$ represents the Euclidean distance between input vector (x) and the radial basis function center (c_j). While r_j is the width of radial basis function.

The output layer operation is linear, given by equation 7:

$$y(x) = \sum_j^n w_j \cdot \Phi_j(x) \tag{7}$$

where w_j are the connection weight of hidden layer to output layer and n is number of hidden node.

It is obvious that RBFNN is simply linear combination and to find the easiest solution linear optimization method is used. It converges for a short time and it guaranties convergence to global optimum parameter. Once applying the training process parameters should be defined as follows: a. Number of nodes in hidden layers, b. the width and center of each radial basis function in each node and c. connecting weight of hidden layer with output layer.

In order to determine the center and the optimum number of hidden nodes learning algorithm based on orthogonal least square (OLS) is applied [17]. This algorithm function in a forward selection manner selecting the center of a radial basis function one by one in rational way until construction of a suitable network.

3. Discussion results

After application of above two techniques it was possible to get results and to analyse them.

3.1 Multiple regression analysis

The analysis starts from description of variables (see Table 2) which take part in setting up the model following with correlative analysis of them and getting general statistical data (see Table 3) as well as histogram for a significant model (see Fig.2). Some models are important but we have selected like a final model which give us best results in forecasting abilities (see Table 3). The best model of selection is enabled after use of stepwise procedure in SPSS software [11].

Table 2: Descriptive of variables

Productive	Name of variable	Mean	Min	Max	StD.	N
$Y_{Productive}$	Traffic_LOG	2.1030	0.48	4.07	0.9730	39
X_{1D}	Rezidents	47164	1724	204721	45479	39
X_{2D}	ActiveForces	11087	485	58708	11883	39
X_{3D}	Area	298	11	633	176	39
X_{4D}	ActiveBussines	1084	31	10182	1757	39
X_{5D}	Pass.Vehicles	7516	73	48381	8606	39
X_{6D}	Inco. per capita	9482	2242	17772	3132	39
X_{7D}	Density	159	20	531	111	39
X_8	Time	70	5	136	38	39
X_9	Distance	69	10	141	39	39
X_{10}	TripCost	7	1	14	4	39

Table 3: Model summary of statistical parameters

Model Summary ^a					
Model	R	R ²	Adjusted R ²	Std. Error	D.Watson
Productive	0.893	0.797	0.786	0.450	1.676
ANOVA					
Model	SumSquares	df	Mean Square	F	Sig.(p<0.05)
Regression	28.690	2	14.345	70.82	.000
Residual	7.291	36	0.203		
Total	35.982	38			
Coefficients ^a					
	B	Std.Err	t	VIF	Sig.(p<0.05)
Constant	3.167	.168	18.895		.000
Act.ForcesD	3.366E-05	.000	5.479	1.00	.000
Time	-0.021	.002	-10.601	1.000	.000

a. Dependent Variable: ProductiveTrip_LOG

Significant model for productive traffic is given by equation 8:

$$Y_{Productive} = 1468.92 \cdot X_{2D}^{(3.66E-05)} / X_8^{(0.021)} \tag{8}$$

All of the variables are significant at the 5% significance level (95% confidence level) for this model. In other words, (P-value) is <0.05 for all independent variables. It is proved that this is the best for all models because it has the best R², and the lowest ME and

RMSE compared to all other models with the combinations of other variables.

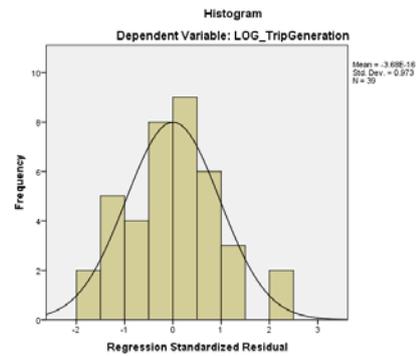


Fig 2. Histogram of regression significant model

The histogram also shows visually that there is approximately normal distribution. Analyzing signs before variables it means that with increase of active forces (X_{2D}) out the region there is increase of traffic while with increase of travel time (X_8) there is decrease of traffic.

3.2 RBFNN analysis

In order to get models through RBFNN we have followed two variants. The first variant: the input variables (13 variables) are in input layer. One hidden layer, and one desired variable (Y) is in output layer with 39 observations are used.

The second variant: only two significant variables which are gained by regression are taken in input layer. One hidden layer, and one desired variable (Y) is in output layer with 39 observations are used. After several attempts it was obvious that the variant with two variables provide much better results of forecasting, respectively in low errors in forecasting. Therefore the further analysis is oriented only to this variant. Data were standardized using the following equation 9:

$$Z = (X - \mu) / \sigma \tag{9}$$

where X is observed value, μ is the mean and σ is the standart deviation. This standartization process yields variable Z with zero mean and unit variance.

For RBFNN analysis, 26 cases (70 %) were assign to the training sample, 9 cases (20%) to the testing sample and 4 cases (10%) to holdout sample. Network information is given in table 4.

Table 4: RBFNN network information

Input Layer	Covariates	1	ActiveForcesD
		2	Time
	Number of Units		2
	Rescaling Method for Covariates		Standardized
Hidden Layer	Number of Units		6 ^a
	Activation Function		Softmax
Output Layer	Dependent Variables	1	TripProduction
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

a. Determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

The RBFNN structure has six units in hidden layers as presented in Fig.3. Also, in table 6 are presented RBFNN model summary. There we can see fewer errors in training than that in testing and holding samples.

Table 5: RBFNN model summary

Training	Sum of Squares Error	0.059
	Relative Error	0.050
	Training Time	0:00:00.02
Testing	Sum of Squares Error	.085 ^a
	Relative Error	.140
Holdout	Relative Error	.427

Dependent Variable: Trip Production

a. The number of hidden units is determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

This is due to the smaller number which is taken into account in testing and holdout samples. However, relative errors in testing and holdout samples are marginal and this has an impact that RBFNN model can be used in the future with high consistency. The architecture and importance of variables of the RBFNN model with two variables in input layer are shown in Fig. 3 and Fig 4.

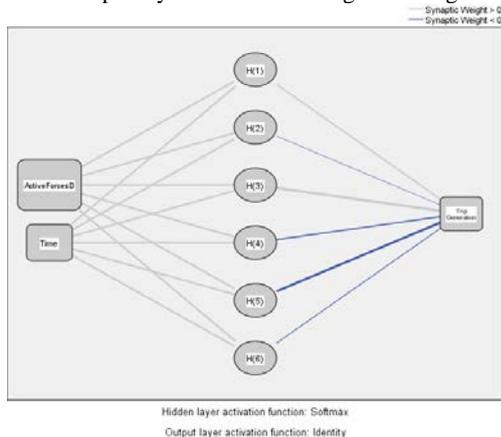


Fig 3.RBFNN Network structure

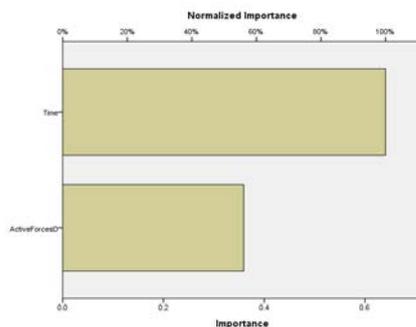


Fig 4. Importance analysis of predictors

3.3 Performance indicators

The most important rule for choosing a forecasting method is its accuracy, or how well they predict matches the actual future value. Here we have used several measures to determine the accuracy of forecasting models and evaluate the performance of models such as: R², ME and RSME.

3.3.1 The coefficient of determination-R²

Coefficient of determination R² can be computed for MRA and RBFNN models using by equation 10:

$$R^2 = \frac{\sum_{i=1}^n [(F_i - \bar{F}) \cdot (A_i - \bar{A})]^2}{[\sum_{i=1}^n (F_i - \bar{F})^2] \cdot [\sum_{i=1}^n (A_i - \bar{A})^2]} \quad (10)$$

where F_i is the ith forecast, A_i is the ith actual values and n-is the number of samples. While \bar{F} and \bar{A} are the averages.

Based on this for regression model we have value of R²=0.797, while for RBFNN (RBF 7-2-1) we have calculated separately for training and for testing by expressions below:

$$R_{training}^2 = 1 - \frac{0.059}{39} = 0.998 \quad R_{testing}^2 = 1 - \frac{0.085}{6} = 0.991$$

Variables were rescaled by standardization, it has zero mean and unit variance, and that sum of squares total is equal n-1, where n is the sample size.

3.3.2 ME and RSME

These indicators can be calculated separately for two models by equation (11) and (12):

$$ME = \frac{\sum_{i=1}^n (F_i - A_i)}{n} \quad (11)$$

$$RSME = \left[\frac{\sum_{i=1}^n (F_i - A_i)^2}{n} \right]^{\frac{1}{2}} \quad (12)$$

Table 6: Summary of indicators of R², ME and RSME

Model	R ²	ME	RSME
Regression model	0.797	-146	1778
RBFNN model	0.998	-10	241

Comparing results of Table 6 show that RBFNN model performs much better than ordinary regression in terms of performance indicators explained by the models.

4. Conclusion

In this paper some models are developed but only two of them are selected which is (MRA) and (RBFNN) to get results in traffic production in Anamorava region. The results of those models have been compared to see differences and advantages provided by each of them in forecasting, respectively in finding out which of them provide smaller errors in forecasting. After verification of indicators such is R², ME and RSME by performance analysis we came to conclusion that RBFNN model provide better performance and as such can be finally used for traffic production forecasting in this region.

5. References

- [1] Elkafoury A. et al. Develop Trip Generation model for Alexandria City-Egypt. Journal of Fluids, Heat and Mass Transfer, Mechanical and Civil Engineering. ISBN:978-1-61804-358-0.
- [2] Arliansyah J. And Haronti Y.: Trip attraction model using radial basis function neural networks. The 5th International Conference of Euro Asia Civil Engineering Forum (EACEF-5). Elsevier & Procedia Engineering, 2015.
- [3] Agency Statistics of Kosovo. General Statistics of Kosovo. Prishtine, 2016.
- [4] Regional Development Strategy for the Economic Region East. Implemented by RDA. Prishtina, 2010.
- [5] Ministry of Infrastructure. Sectorial Strategy of Multimodal Transport 2015-2025 and Action Plan. Prishtina, 2015.
- [6] Ortuzar J.D, Williamson L.G., Third Edition. Modelling Transport.
- [7] Fricker J.D, Saha S.K., 1987. Final Report. Traffic Volume Forecasting Methods for Rural State Highways. Joint Highway Research Project FHWA/IN/JHRP-86/20.
- [8] Semeida A.M., 2014. Derivation of Travel Demand Forecasting Models for Low Population Areas: The case of Port Said Governorate, North East Egypt. Journal of Traffic and Transportation Engineering (English Edition). 1(3):196-2008,
- [9] Varagouli E.G, Simos T.E, Xeidakis G.S., 2005. Fitting a Multiple Regression Line to Travel Demand Forecasting: the Case of Prefecture of Xanthi, Northern Greece. Mathematical and Computer Modelling, 42 (7/8):817-836.
- [10] Navidi W., Statistics for Engineers and Scientists. McGrawHill-Higher Education. ISBN-13:978-0-07-255160-0 and ISBN-10:0-07-255160-7.
- [11] Kalayaci S., 2014.SPSS in Practice. Statistical Techniques with Many Variables. Six Edition.
- [12] Anderson D.R, Sweeney D.J, Williams Th.A., Statistics for Business and Economy. Second piece. Translated in Albanian by published company "UET Press. Tirana.
- [13] Saha K.Sunil and Fricker J.D.: The development of a procedure to forecast traffic volumes on urban segments of the state and intersate highway systems., Final report FHWA/IN/HJHP-90/11, Joint Highway Research Project. School of Civil Engineering, Purdue University, 1990.
- [14] Cicak M.: Modeliranje u zeleznickom saobracaju. Saobracajni Fakultet-Univerziteta u (Book in Serbian). Beogradu,2003.
- [15] Stamenkovic L.J.: Predvidanje emisije gasovnih zagadujucih materija na nacionalnom nivou primenom modela zasnovanih na vesticim neuronskim mrezaama. Doktorska disertacija. Teholoski-Matalurski Fakultet, Universitet u Beogradu, 2016.
- [16] Aguilar D.P.: A radial basis neural network for the analysis of transportation data. Graduate thesis. University of South Florida, 2004.
- [17] Chen, S., Cowan, C. F. N. and Grant, P. M., Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks, IEEE, Transaction on Neural Networks, Vol. 2, No. 2, 302-309, 1991.