**FACTA UNIVERSITATIS**
Series: **Mechanical Engineering** Vol. 13, N$^{o}$ 2, 2015, pp. 133 - 141

**Original scientific paper**

# RAIL TRAFFIC VOLUME ESTIMATION BASED ON WORLD DEVELOPMENT INDICATORS

*UDC 656.2*

**Luka Lazarević, Miloš Kovačević, Zdenka Popović**

Faculty of Civil Engineering, University of Belgrade, Serbia

**Abstract**. *European transport policy, defined in the White Paper, supports shift from road to rail and waterborne transport. The hypothesis of the paper is that changes in the economic environment influence rail traffic volume. Therefore, a model for prediction of rail traffic volume applied in different economic contexts could be a valuable tool for the transport planners. The model was built using common Machine Learning techniques that learn from the past experience. In the model preparation, world development indicators defined by the World Bank were used as input parameters.*

**Key words**: *rail traffic, prediction, Machine Learning, World Bank, development indicators*

## 1. INTRODUCTION

Traffic volume prediction is very important task from the planner`s point of view. It can be used for different activities such as Transport market survey and analysis, assessment of the transport market demands and level of provided services, development of appropriate plans, measures and strategies, and decision about new investments in infrastructure.

There are several different approaches for the traffic prediction, depending on the planner needs. Regarding the time interval, it can be performed as a short-term [1-4] or a long-term prediction [5-7]. In addition, the prediction could be performed on the level of state transport network [4-6, 8-10], urban transport network [2, 3, 7, 11, 12], or particular section in the transport network [1].

Previous research showed that the application of neural networks for traffic prediction problems provided good results [2, 3, 9, 11, 12]. Neural networks (NN) exhibited the flexibility in modeling complex datasets with possible nonlinearities or missing data [13]. Further, Support Vector Machines (SVM) provided good prediction results in some cases [6, 7].

Received January 30, 2015 / Accepted April 15, 2015
**Corresponding author**: Luka Lazarević
Faculty of Civil Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, Belgrade, Serbia
E-mail: llazarevic@grf.bg.ac.rs

Qi et al. used neural tree model for prediction of railway passenger traffic. This model provided better results than NN and SVM [10]. The similar approach was applied in the research by Zhuo et al [9]. This research applied back propagation NN to predict railway passenger volume resulting in quick convergence and high accuracy. Both studies implied time-series analysis.

Li et al. proposed factors of urban rail transit flow, which were used in the prediction of Shanghai Metro traffic [7]. Research by Griskeviciene et al. [5] proposed key macroeconomic indicators that are related to railway traffic. The prediction model was developed using previous statistical data, and applied in three prospective scenarios - optimistic, basic (realistic) and pessimistic, for prediction of long-term freight traffic on railway corridors IX and I in Lithuania.

The aim of this paper is creation of a data driven model for prediction of changes in rail traffic volume based on the changes in economic parameters. World development indicators defined by the World Bank [14] were used as inputs to the proposed models. Two countries were chosen for the creation of the model, Serbia and Austria, since they have similar area, population and population density [15]. The freight and passenger traffic were separately analysed for both countries.

The models were developed using Weka 3 software, a collection of machine learning algorithms for data mining tasks, developed at the University of Waikato [16, 17].

## 2. DATA REPRESENTATION

The first task was to choose the appropriate data representation and prepare a dataset that will be used for building and validation of the proposed prediction models. As it was stated above, world development indicators were chosen as economic parameters to represent the predictors of the traffic for each year in both countries. World Bank (WB) defines 1356 indicators divided into ten groups: *education (103)*, *environment (136)*, *economic policy and debt (506)*, *financial sector (50)*, *health (122)*, *infrastructure (36)*, *social protection and labor (148)*, *poverty (22)*, *private sector and trade (151)*, and *public sector* (82). More details about these indicators can be found on the WB website (http://data.worldbank.org). Two of the infrastructure indicators represent data about freight and passenger rail traffic (target values to predict). WB databank provides rail traffic data from 1980 to 2012. Therefore this time range was chosen to build the proposed models.

Since the indicators have different orders of magnitude and measure units, their values were replaced with relative changes (changes in percentage comparing to the previous year). Modifications on the initial dataset enabled better analysis of the correlation between economic changes and changes in rail traffic.

From the aspect of Machine Learning, relative changes of world development indicators represent numerical attributes. On the other hand, relative changes of rail traffic were represented as nominal values (classes):

- decrease in rail traffic was denoted as N (relative change is negative), and
- increase in rail traffic was denoted as P (relative change is positive).

Introduction of two classes transformed the traffic volume prediction problem into a simple classification task where the change in rail traffic is classified as positive or negative based on the relative changes of world development indicators in the related country.

The next step in data preparation was to eliminate the attributes with more than 50% of missing values. The number of attributes (*n*) was not the same for Serbia and Austria, since it depends on availability of data. Fig. 1 shows the data representation used to build the initial dataset for each country.

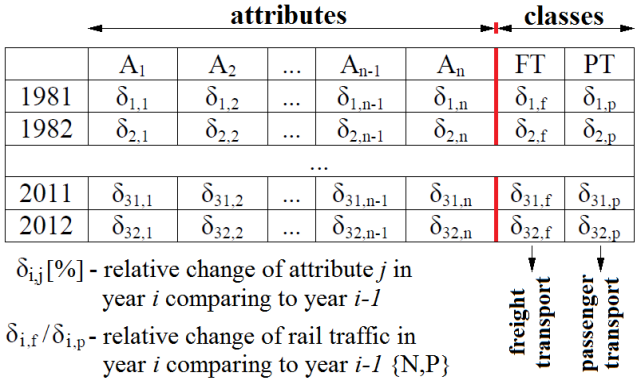


**Fig. 1** The proposed data representation

A separate set was created for each traffic type from the country's initial dataset by retaining the appropriate class attribute (freight or passenger) resulting in four datasets $D_i$, $D_1$ = SerbianFreight (SF), $D_2$ = SerbianPassenger (SP), $D_3$ = AustrianFreight (AF) and $D_4$ = AustrianPassenger (AP). Although these datasets did not contain same number of attributes for both countries, they were used for assessing the performances of prediction models that were initially built.

## 3. Building and Validating the Models

In order to build and validate the model using a Machine Learning approach, dataset $D_i$ should be divided into disjoint training and test sets. Since $D_i$ contained small number of examples a valid statistical protocol for model validation demanded the creation of *k* disjoint train-test splits. The idea of this procedure is to train *k* different models tested on *k* different test sets and to average the obtained performances. In this paper five train-test splits were created for each $D_i$ containing 80% of examples for training and 20% for testing.

The next problem to be solved considered the fact that the number of attributes (world development indicators) in each of the four sets was significantly higher than the number of examples, or $j >> i$ according to Fig. 1. Therefore, the number of attributes was reduced using the correlation feature subset (CFS) filter. This filter selects a subset of attributes that are highly correlated with the class (N or P in this case), while having low inter-correlation. The filter was applied on each of the five train-test splits, for each $D_i$.

In this study we applied three commonly used Machine Learning classifiers to build the models with training data for each $D_i$:

- Naive Bayes (NB) – a probabilistic classifier based on Bayes theorem and the assumption of mutual independence of attributes [18],
- Decision Tree (DT) – a classifier which successively tests attribute values in each internal node until it is possible to deduce about the item`s target value (class) [19],
- Multilayer Perceptron (MLP) – feed-forward neural network that maps sets of input data onto a set of appropriate outputs, trained using back-propagation algorithm [20].

After applying classifiers on each train-test split five confusion matrices were obtained. A confusion matrix $\mathbf{C}_{n\,x\,n}$ ($n$ is the number of classes) is defined with $c_{ij}$ representing the number of cases from actual class $i$ that are classified as class $j$. When generating train-test splits, we applied cross-validation protocol which ensured that the test parts were mutually disjoint. Therefore, the final confusion matrix for the model was obtained by simple addition of separate matrices. Since the related classification problem assumed the existence of only two classes, the final confusion matrix has the form given with Eq. (1):

$$\begin{bmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{bmatrix} \tag{1}$$

where: TP - the number of instances correctly classified as P,
TN - the number of instances correctly classified as N,
FN - the number of instances incorrectly classified as N,
FP - the number of instances incorrectly classified as P.

Using data from the confusion matrix, three pieces of information related to both classes (P, N) were derived for each model: precision ($\pi$), recall ($\rho$) and F-measure (Eq. 2-5).

$$\pi_{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad \pi_{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \tag{2}$$

$$\rho_{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \rho_{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

$$\mathrm{F}_{\mathrm{P}} = 2\cdot\frac{\rho_{\mathrm{P}}\cdot\pi_{\mathrm{P}}}{\rho_{\mathrm{P}}+\pi_{\mathrm{P}}} \qquad \mathrm{F}_{\mathrm{N}} = 2\cdot\frac{\rho_{\mathrm{N}}\cdot\pi_{\mathrm{N}}}{\rho_{\mathrm{N}}+\pi_{\mathrm{N}}} \tag{4}$$

$$\mathrm{F} = \frac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}\mathrm{F}_{\mathrm{P}} + \frac{\mathrm{TN}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}\mathrm{F}_{\mathrm{N}} \tag{5}$$

Class precision measures the percentage of correct decisions among all decisions for the specified class while the class recall measures the percentage of recognized cases from the class. Increasing the precision often leads to decreasing in recall. F-measure represents the harmonic mean between the two describing the overall class performance of the model. In order to compare different models with only one measure, a weighted F is defined with Eq. 5. The obtained performances of all models are presented in Table 1.

**Table 1** Results of the prediction models (best cases underlined)

| Classifier | Model | $F_P$ | $F_N$ | F |
|---|---|---|---|---|
| Naive Bayes (NB) | SF | 0.632 | 0.462 | 0.558 |
| | SP | 0.560 | 0.718 | 0.659 |
| | AF | 0.696 | 0.222 | 0.563 |
| | AP | 0.766 | 0.353 | 0.624 |
| Decision Tree (DT) | SF | 0.529 | 0.467 | 0.502 |
| | SP | 0.500 | 0.700 | 0.625 |
| | AF | 0.708 | 0.125 | 0.544 |
| | AP | 0.632 | 0.462 | 0.568 |
| Multilayer Perceptron (MLP) | SF | 0.579 | 0.385 | 0.494 |
| | SP | 0.583 | 0.750 | 0.687 |
| | AF | 0.571 | 0.182 | 0.462 |
| | AP | 0.732 | 0.522 | 0.653 |

SF, SP - freight and passenger traffic in Serbia
AF, AP - freight and passenger traffic in Austria

Weighted F-measure of the developed models ranged from 0.56 to 0.69, depending on the country and traffic type. The average performance of the models could be explained with the small dataset and the unequal number of examples per class, except in the case of freight traffic in Serbia where the class split was almost 50%.

According to the values from Table 1, NB provided better results for freight traffic and MLP provided better results for passenger traffic. In addition, predicting the passenger traffic for both countries appeared to be the easier task than predicting the freight traffic. Predicting decrease in freight traffic in Austria was the most difficult task according to the low $F_N$ value.

## 4. Influence of the World Development Indicators

As it was mentioned before, number of attributes (world development indicators) in training sets from each $D_i$ was reduced using the CFS filter. According to the definition of CFS filter, there were actually created subsets of world development indicators that are mostly correlated with the rail traffic changes.

After the analysis of selected WD indicators, it was noted that rail traffic mostly depends on indicators belonging to the two groups: *environment* and *economic policy and debt.* Fig. 2 shows the distribution of selected WD indicators over the four significant groups.

For example, from the *environment group*, parameters related to the $CO_2$ emission can be considered as informative attributes (they often repeated in data subsets). In particular, it was determined that the increase of rail traffic in Austria was followed by decrease of $CO_2$ emission from liquid fuel consumption and $CO_2$ intensity (in 70% of examples). Although this complies with the expected effects of the shift to rail transport, this type of analysis should also consider many other aspects of economy.

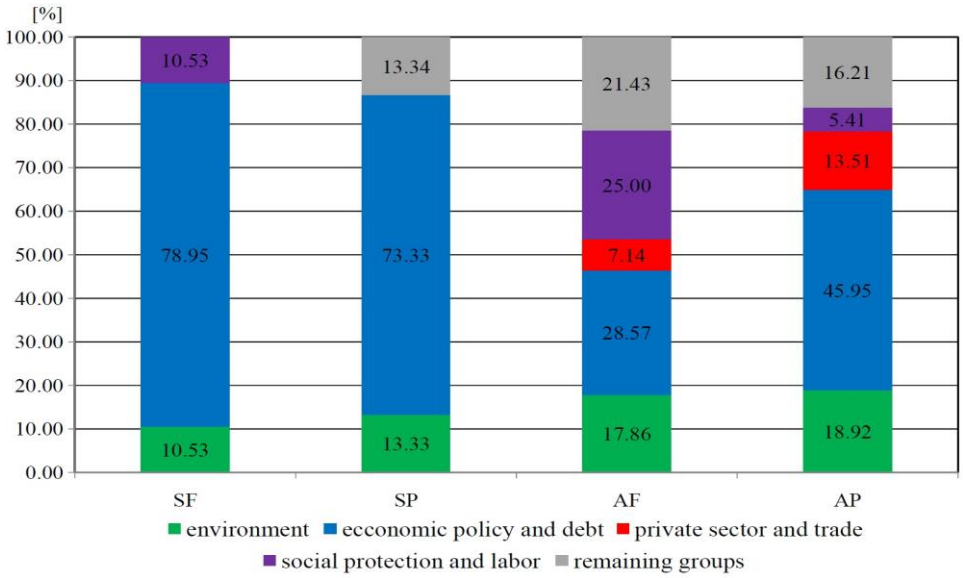


**Fig. 2** Distribution of selected WD indicators over the four significant groups

From the *economic policy and debt group*, gross national income and gross national expenditure appeared to be informative.

In addition, several WD indicators could not be set in the context of the model, although they were selected by the CFS filter. For example, several indicators originated from the health and education groups.

However, application of CFS filter showed which groups of indicators are mostly correlated with rail traffic and which indicators can be used for traffic prediction.

Among the indicators that were selected by the CFS filter, the ones that repeated in all sets (according to Fig. 3) were used for preparation of training and test sets according to the methodology described in Section 2. Table 2 presents indicators that were chosen for the prediction model.

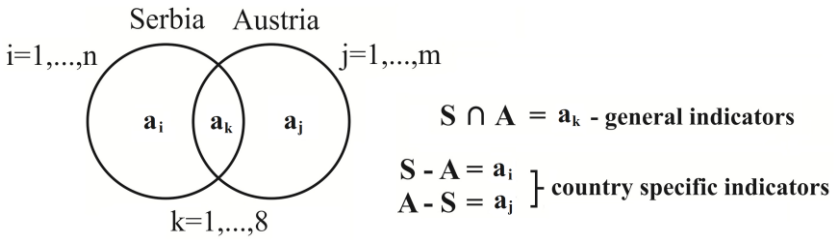


**Fig. 3** Selection of the relevant indicators

**Table 2** Chosen indicators for the prediction models

| Attribute (indicator) |
|---|
| $CO_2$ emissions from transport (million metric tons) |
| Energy production (kt of oil equivalent) |
| Energy use (kg of oil equivalent per capita) |
| Adjusted savings: energy depletion (current US$) |
| GDP per capita (current international $) |
| GNI per capita (current international $) |
| Net income from abroad (current US$) |
| Road sector energy consumption (kt of oil equivalent) |

After the models were recreated using the attributes from Table 2 the obtained performances are presented in Table 3.

**Table 3** Final results of the prediction models (best cases underlined)

| Classifier | Model | $F_P$ | $F_N$ | F |
|---|---|---|---|---|
| Naive Bayes (NB) | SF | 0.700 | 0.500 | 0.613 |
| | SP | 0.240 | 0.513 | 0.411 |
| | AF | 0.784 | 0.154 | 0.607 |
| | AP | 0.667 | 0.571 | 0.631 |
| Decision Tree (DT) | SF | 0.698 | 0.381 | 0.559 |
| | SP | - | 0.638 | - |
| | AF | 0.766 | 0.353 | 0.650 |
| | AP | 0.718 | 0.560 | 0.659 |
| Multilayer Perceptron (MP) | SF | 0.500 | 0.500 | 0.500 |
| | SP | 0.320 | 0.564 | 0.473 |
| | AF | 0.711 | 0.316 | 0.600 |
| | AP | 0.571 | 0.182 | 0.425 |

SF , SP - freight and passenger traffic in Serbia
AF , AP - freight and passenger traffic in Austria

NB provided good result in predicting the increase of freight traffic in Serbia. Although there was almost equal number of examples from two classes, F-measure for prediction of freight traffic decrease was only 0.5. On the other hand, prediction of passenger rail traffic in Serbia was more difficult task. DT provided best result for prediction of traffic decrease, but failed to predict the traffic increase (class P). The main reason for this result is unequal class split in dataset for passenger traffic (72.5% of class P and 37.5% of class N) and large number of missing values for the used attributes.

DT provided good results for both traffic types in Austria. Prediction of freight traffic decrease had significantly low performance due the small number of examples for class N in dataset (about 28%).

Comparing to Table 1, better results were obtained for freight traffic in both countries. Result for passenger traffic in Austria was not improved, while in case of Serbia result

was significantly lower. Therefore, attributes presented in Table 2 do not reflect the changes in passenger rail traffic in Serbia.

It is important to mention that the number of missing values ranged from 35-50% of all values depending on the attribute and country. Therefore, obtained results were highly influenced by the large number of missing values.

## 5. DISCUSSION AND CONCLUSION

Countries with strong industry and economy have fully organized and functional railway transport. The main reason is high capacity, efficiency and safety of railways. Hence, European transport policies are directed towards the shift to rail transport.

The shift to rail transport will be followed by certain economic changes. On the other hand, economic changes can influence the changes in rail traffic volume in countries that strongly rely on rail transport.

The research presented in this paper was directed towards the development of rail traffic prediction models based on Machine Learning techniques which utilize world development indicators defined by the World Bank. Models were developed and evaluated for two countries, Serbia and Austria, in order to provide sound basis for model comparisons.

Performances of the prediction models were assessed using F-measure. Considering that minimum required F-measure should be 0.75, obtained performances for all models were below this threshold. However, prediction models provided good estimation of changes in rail traffic regardless of the small dataset and large number of missing values. Therefore, these models can be used for preliminary estimations for the purposes of transport market survey and analysis, as well as for development of plans, measures and strategies on the level of network or its sections.

Further research in this field is directed towards determination of the general indicators using several different countries, which would provide larger dataset and thus more advanced prediction model. The main goal would be to create the general model for railway traffic prediction based on world development indicators.

**Acknowledgement**: *This paper was supported by the Ministry of Science and Technological Development of the Republic of Serbia through the research project No. 36012: „Research of technical-technological, staff and organizational capacity of Serbian Railways, from the viewpoint of current and future EU requirements".*

## REFERENCES

1. Abdulhai B., Porwal H., Recker W., 2002, *Short-Term Traffic Flow Prediction Using Neuro-Genetic Algorithms*, Journal of Intelligent Transportation Systems, 7(1), pp. 3-41.
2. Celikoglua H.B., Cigizoglu H.K., 2007, *Public transportation trip flow modeling with generalized regression neural networks*, Advances in Engineering Software, 38(2), pp. 71–79.
3. Çetiner B.G., Sari M., Borat O., 2010, *A neural network based traffic-flow prediction model*, Mathematical and Computational Applications, 15(2), pp. 269-278.
4. Guo F., Krishnan R., Polak J., 2013, *A computationally efficient two-stage method for short-term traffic prediction on urban roads*, Transportation Planning and Technology, 36(1), pp. 62-75.

5. Griskeviciene D., Griskevicius A., Griskeviciute-Geciene A., 2010, *Providences and projections regarding the prognostication of railway transport volumes from a long-term perspective*, Proc. Tenth International Conference "Reliability and Statistics in Transportation and Communication", Riga, Latvia, pp. 25-33.
6. Gao S., Zhang Z., Cao C., 2011, *Road Traffic Freight Volume Forecast Using Support Vector Machine Combining Forecasting*, Journal of Software, 6(9), pp. 1680-1687.
7. Li Z., Zhang Q., Wang L., 2011, *Flow prediction research of urban rail transit based on support vector machine*, Proc. First International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, pp. 2276-2282.
8. Fang H., Yaqiang S., Siyu T., 2007, *The application of combined forecast method in predicting freight volume of railway*, Proc. First International Conference on Transportation Engineering, Southwest Jiaotong University, Chengdu, China, pp. 3347-3352.
9. Zhuo W., Li-Min J., Yong Q., Yan-Hui W., 2007, *Railway passenger traffic volume prediction based on neural network*, Applied Artificial Intelligence, 21(1) , pp. 1-10.
10. Qi F., Liu X., Ma Y., 2009, *Prediction of Railway Passenger Traffic Volume Based on Neural Tree Model*, Proc. Second International Conference on Intelligent Computation Technology and Automation, Washington, USA, pp. 370-373.
11. Celikoglua H.B., Cigizoglu H.K., 2007, *Modelling public transport trips by radial basis function neural networks*, Mathematical and Computer Modelling, 45(3-4), pp. 480-489.
12. Özuysal M., Tayfur G., Tanyel S., 2012, *Passenger flows estimation of light rail transit (LRT) system in Izmir, Turkey using multiple regression and ANN methods*, Promet - Traffic&Transportation, 24(1), pp. 1-14.
13. Karlaftis M.G., Vlahogianni E.I., 2011, *Statistical methods versus neural networks in transportation research: Differences, similarities and some insights*, Transportation Research Part C: Emerging Technologies, 19(3), pp. 387-399.
14. http://data.worldbank.org/ (Accessed on December 26, 2015)
15. Popović Z., Lazarević L., Iždvolt L., 2013, *Potential of the railway infrastructure in Serbia*, Railway transport and logistics, 3, pp. 9-22.
16. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann M., Witten I.H., 2009, *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, 1(11), pp. 10-18.
17. http://www.cs.waikato.ac.nz/ml/weka/ (Accessed on December 26, 2015)
18. Mitchel T., 1997, *Machine Learning*, McGraw Hill, Columbus, Ohio, 414 p.
19. Quinlan J.R., 1986, *Induction of Decision Trees*, Machine Learning, 1(1), pp. 81-106.
20. Haykin S., 1998, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New Jersey, 842 p.