University of Niš
Faculty of Mechanical Engineering

PROCEEDINGS

XVI SCIENTIFIC-EXPERT CONFERENCE ON RAILWAYS
RAILCON '14

October 09 - 10, 2014, Niš, Serbia

Publisher

Faculty of Mechanical Engineering Niš
Prof. Vlastimir Nikolić, Dean

Editor

Prof. Dušan Stamenković
Prof. Miloš Milošević

Reviewers

Prof. Dušan Stamenković
Prof. Miloš Milošević
Prof. Zdenka Popović

Technical preparation

Milan Banić
Aleksandar Miltenović
Nikola Petrović

Cover design

Prof. Miloš Milošević

Number of copies

160

Printing

UNIGRAF, Niš

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

629.4(082)
656.2(082)
625.1(082)
338.47(497.11)(082)

SCIENTIFIC-Expert Conference of Railways (16th ; 2014 ; Niš)
Proceedings / XVI Scientific-Expert Conference on Railways - RAILCON '14, October 09-10, 2014, Niš ; [organized by] University of Niš, Faculty of Mechanical Engineering ; [editor Dušan Stamenković]. - Niš : Faculty of Mechanical Engineering, 2014 (Niš : Unigraf). - XXI, 282 str. : ilustr. ; 25 cm

Tiraž 160. - Napomene uz tekst. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-6055-060-8

1. Faculty of Mechanical Engineering (Niš)
a) Железничка возила - Зборници b) Железнички саобраћај - Зборници c) Железничке пруге - Зборници d) Србија - Саобраћајна политика

COBISS.SR-ID 210153228

ISBN 978-86-6055-060-8

Ministry of Education, Science and Technological Development of the Republic of Serbia has participated in printing costs of the Proceedings of the XVI International Scientific-expert Conference on Railways - **RAILCON 2014**

All the publications in these Proceedings have the authorship, whereas the authors of the papers carry entire responsibility for originality and content.

University of Niš

XVI SCIENTIFIC-EXPERT CONFERENCE ON RAILWAYS
RAILCON '14

Faculty of Mechanical Engineering

Program Committee

Organizing Committee

Patrons

Ministry of Education, Science and Technological Development of the Republic of Serbia
JSC SERBIAN RAILWAYS
MIN Lokomotiva a.d.

Cofinanciers

STADLER
ALTPRO
STRAIL
MIN DIV Svrljig
THALES
OPEN TRACK

XVI SCIENTIFIC-EXPERT CONFERENCE ON RAILWAYS
RAILCON '14

DATA DRIVEN MODEL FOR PREDICTION OF RAIL TRAFFIC

Luka LAZAREVIĆ [1]
Miloš KOVAČEVIĆ [2]
Zdenka POPOVIĆ [3]

Abstract *– The rail system, as a system with high capacity, the least air and water pollution, solvable noise and vibration emission problem and the least space usage, is competitive with other modes of transportation. In addition, European transport policy supports shift from road to rail and waterborne transport. It is expected that changes in rail traffic volume be followed by changes in certain economic parameters. On the other hand, changes in economy influence rail traffic volume. The aim of this paper is creation of the data driven models for prediction of rail traffic volume in different economic contexts, using world development indicators, defined by the World Bank, as input parameters.*

Keywords – Rail traffic, data driven model, prediction, World Bank, development indicators.

1. INTRODUCTION

European transport policies were at first directed at the development of the road network. Fast increase in number of vehicles resulted in higher road congestion, and thus the increase of air pollution and the number of traffic accidents. It was obvious that road development has limited potential, especially in densely populated areas.

Although construction costs for roads are, in general, lower than construction costs for railways, for high traffic volumes external cots are significantly lower on railways. Therefore, the rail system as a high capacity system with the least air and water pollution, solvable noise and vibration emission problem and the least space usage became competitive with other modes of transportation.

The European Union (EU) defined a common transportation policy in its White Paper entitled "Roadmap to a Single European Transport Area - Towards a competitive and resource efficient transport system" [1]. This policy defines 10 goals for reaching a competitive and resource efficient transport system. Four of these goals are directed towards the use of rail transport:

- to shift 30% of road freight over 300 km to rail or waterborne transport by 2030, and more than 50% by 2050,
- to complete European high-speed rail network by 2050, triple the length of the existing high-speed rail network by 2030 and maintain a dense railway network in all Member States,
- to complete a fully functional and EU-wide multimodal Trans-European Transport Network (TEN-T) by 2030, with a high quality and capacity network by 2050 and a corresponding set of information services, and
- to connect by 2050 all network airports to the rail network, preferably high-speed, and to ensure that all seaports are sufficiently connected to the rail freight and, where possible, inland waterway system [1].

It is expected that achievement of these goals will be followed by certain economic changes. On the other hand, economic changes can influence the changes in rail traffic volume.

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. Two countries were chosen for the creation of the model, Serbia and Austria since they have similar area, population and population density [2]. For both countries, it was separately analysed freight and passenger traffic.

As economic parameters, there were used world development indicators defined by the World Bank [3]. The models were developed using software Weka 3, a collection of machine learning algorithms for data mining tasks, developed at the University of Waikato [4,5].

[1] Luka LAZAREVIĆ, master inž.građ., University of Belgrade, Faculty of Civil Engineering, llazarevic@grf.bg.ac.rs
[2] Miloš KOVAČEVIĆ, dipl.inž.elektr., University of Belgrade, Faculty of Civil Engineering, milos@grf.bg.ac.rs
[3] Zdenka POPOVIĆ, dipl.građ.inž., University of Belgrade, Faculty of Civil Engineering, zdenka@grf.bg.ac.rs

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 model. 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) divides these indicators into ten groups: education, environment, economic policy and debt, financial sector, health, infrastructure, social protection and labor, poverty, private sector and trade, and public sector [3]. Two of the infrastructure indicators are data about freight and passenger rail traffic. 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). The same processing was applied on the rail traffic data. Both modifications on the dataset ensured the model aim - establishing 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 problem of prediction of rail traffic into a simple classification task. The prediction model was built from the data to classify the change in rail traffic whether as positive or negative based on the relative changes of world development indicators in the related country.

attributes | classes

	A_1	A_2	...	A_{n-1}	A_n	FT	PT
1981	$\delta_{1,1}$	$\delta_{1,2}$	...	$\delta_{1,n-1}$	$\delta_{1,n}$	$\delta_{1,f}$	$\delta_{1,p}$
1982	$\delta_{2,1}$	$\delta_{2,2}$	...	$\delta_{2,n-1}$	$\delta_{2,n}$	$\delta_{2,f}$	$\delta_{2,p}$
				...			
2011	$\delta_{31,1}$	$\delta_{31,2}$	...	$\delta_{31,n-1}$	$\delta_{31,n}$	$\delta_{31,f}$	$\delta_{31,p}$
2012	$\delta_{32,1}$	$\delta_{32,2}$	...	$\delta_{32,n-1}$	$\delta_{32,n}$	$\delta_{32,f}$	$\delta_{32,p}$

$\delta_{i,j}$ [%] - relative change of attribute *j* in year *i* comparing to year *i-1*

$\delta_{i,f}/\delta_{i,p}$ - relative change of rail traffic in year *i* comparing to year *i-1* {N,P}

freight transport | passenger transport

Fig.1. The proposed data representation

The last step in data preparation was to eliminate attributes that have more than 50% of missing values. Therefore, 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 (dataset) used to build the model.

3. PREPARATION OF TRAINING AND TEST SETS

After the representation is defined and the entire data set is prepared, it is necessary to divide it into disjoint subsets for building (training set) and validating (test set) the model.

Since it is expected that freight and passenger transport do not depend on the same indicators, separate models were developed for these two traffic types.

Using the software-implemented filter in Weka, the dataset was divided into five non-overlapping folds for the purpose of cross validation. This filter divided dataset on training and test sets in five different ways (each train-test split consisted of 80% of data for training and 20% of data for testing the model).

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

4. MODEL VALIDATION AND OBTAINED RESULTS

According to the previous, four groups of data each containing five train-test splits were created to build and validate four models: for freight and passenger traffic in Serbia, and for freight and passenger traffic in Austria. After learning the mapping from attributes to classes using common machine learning techniques, three classifiers were built for each model:

- Naive Bayes (NB) - a probabilistic classifier based on Bayes theorem and the assumption of mutual independence of attributes [6],
- 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) [7],
- Multilayer Perceptron (MLP) - feed-forward neural network that maps sets of input data onto a set of appropriate outputs, trained using back-propagation algorithm [8].

After applying classifiers on each train-test split generated in a 5-fold cross validation procedure, five confusion matrices were obtained. Confusion matrix for the model is obtained by simple addition of separate matrices:

$$\begin{bmatrix} \sum_{i=1}^{k} TP_i & \sum_{i=1}^{k} FN_i \\ \sum_{i=1}^{k} FP_i & \sum_{i=1}^{k} TN_i \end{bmatrix} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{1}$$

where k is number of folds, and:

- TP (true positive) - the number of instances that are correctly classified as positive,
- TN (true negative) - the number of instances that are correctly classified as negative,
- FN (false negative) - the number of instances that are incorrectly classified as negative, and
- FP (false positive) - the number of instances that are incorrectly classified as positive.

Using data from the confusion matrix, three pieces of information related to both classes (P, N) were derived for each model: precision (π), recall (ρ) and F-measure.

$$\pi_P = \frac{TP}{TP+FP} \qquad \pi_N = \frac{TN}{TN+FN} \tag{2}$$

$$\rho_P = \frac{TP}{TP+FN} \qquad \rho_N = \frac{TN}{TN+FP} \tag{3}$$

$$F_P = 2 \cdot \frac{\rho_P \cdot \pi_P}{\rho_P + \pi_P} \qquad F_N = 2 \cdot \frac{\rho_N \cdot \pi_N}{\rho_N + \pi_N} \tag{4}$$

$$F = \frac{TP+FN}{TP+TN+FN+FP} F_P + \frac{TN+FP}{TP+TN+FN+FP} F_N \tag{5}$$

All the obtained model results using equations (2)-(5) are presented in Tab.1. Models that provided the largest F-measure are marked with gray colour in Tab.1.

Tab. 1. Final results of the prediction models

Classifier	Model	F_P	F_N	F
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
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
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				

Micro weighted F-measure of the developed models ranged from 0.56 to 0.69, depending on the country and traffic type. The main reasons for this performance are:

- small dataset,
- unequal number of examples of one class comparing to other, except in the case of freight traffic in Serbia where class split was almost 50%, and
- the fact that each train-test split consisted of 80% of data for training and 20% of data for testing.

According to the values from Tab. 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 the decrease of freight traffic in Austria was the most difficult task according to the low F_N value.

5. INFLUENCE OF THE WORLD DEVELOPMENT INDICATORS

As it was mentioned before, number of attributes (world development indicators) in the dataset 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.

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.

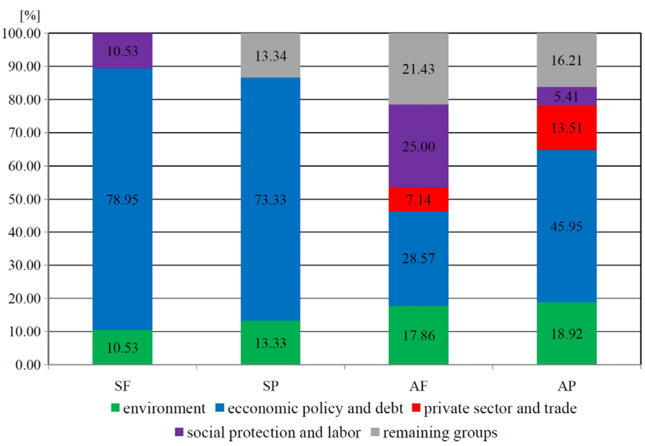

Fig.2. Distribution of selected WD indicators over the four significant groups

For example, from the environment group, as informative indicators it can be considered parameters related to the CO_2 emission and water pollution (that often repeated in data subsets). In particular, it was determined that the increase of rail traffic in Autria 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.

From the economic policy and debt group, as informative indicators it can be considered gross national income and gross national expenditure.

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 prediction of rail traffic.

6. CONCLUSION

In general, it is considered that countries with strong industry and economy have fully organized and functional railway transport. The main reason is high capacity, efficiency and safety of railways. That is why 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 development of the model for prediction of changes in rail traffic based on changes of world development indicators defined by the World Bank. Models were developed for two countries, Serbia and Austria, in order to provide sound basis for model comparisons.

Relevant indicators were chosen according to their correlation with traffic changes. In most cases, selected relevant indicators were CO_2 emission, water pollution, gross national income and gross national expenditure.

Micro weighted F-measure of the developed models ranged from 0.56 to 0.69, depending on the country and traffic type. The main reason for such performance is a small dataset and large difference between the number of positive examples (increase of rail traffic) and the number of negative examples (decrease of rail traffic). However, the research proved that changes in rail traffic are followed by changes in many aspects of country`s economy.

ACKNOWLEDGEMENT

This work 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] European Commission, *White Paper - Roadmap to a single European transport area - Towards a competitive and resource efficient transport system*, Commission communication COM (2011) 144, Brussels, 2011.
[2] Popović, Z., Lazarević, L., Ižvolt, L., *Potential of the railway infrastructure in Serbia*, Railway transport and logistics, Vol. 3, 2013.
[3] http://data.worldbank.org/
[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, M., Witten, I.H., *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, Volume 11, Issue 1, 2009.
[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] Mitchel, T., *Machine Learning*, McGraw Hill, Columbus, Ohio, 1997.
[7] Quinlan, J.R., *Induction of Decision Trees*, Kluwer Academic Publishers, Boston, 1986.
[8] Haykin, S., *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New Jersey, 1998.