Environment Protection Engineering

Vol. 44 2018 No. 1

DOI: 10.5277/epe180110

MARIJA BLAGOJEVIĆ[1], MILOŠ PAPIĆ[1], MOMČILO VUJIČIĆ[1], MARKO ŠUĆUROVIĆ[1]

ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTING AIR POLLUTION. CASE STUDY OF THE MORAVICA DISTRICT, SERBIA

An example of artificial neural network model for predicting air pollution has been presented. The research was conducted in Serbia, the Moravica District, on the territory of two municipalities (Lučani and Ivanjica) and the town Čačak. The level of air pollution was classified by a neural network model according to the input data: municipality, site, year, levels of soot, sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and particulate matter. The model was evaluated using a lift chart and a root mean square error (RMSE) has been determined, whose value was 0.0635. A multilayer perceptron has also been created and trained with a back propagation algorithm. The neural network was tested with the data mining extensions (DMX) queries. The results have been obtained for air pollution based on new input data that can be used to predict the level of pollution in future if new measurements are carried out. A web-based application was designed for displaying the results.

1. INTRODUCTION

Air is one of the most important segments of the environment which , with the expansion of industry, has become contaminated with noxious ingredients that are harmful primarily for man and for his environment. Air pollution have thus become a serious environmental problem, especially in the urban areas because of the impact on the physical and mental health [1, 2]. Numerous sources of air pollution depending on the type of pollutants affect human health, especially the health of the most vulnerable parts of the population (pregnant women, children, old and sick people).

The research deals with prediction of air pollution level using neural networks similarly as in the study reported in [3, 4] where the impact of air pollution on health and mortality has been determined. A large number of studies report on the analysis of

[1]University of Kragujevac, Faculty of Technical Sciences Čačak, Svetog Save 65, 32000 Čačak, Serbia, corresponding author M. Blagojević, e-mail address: marija.blagojevic@ftn.kg.ac.rs

air pollution [3–5] and the use of artificial neural networks for these purposes [6– 9]. Adams et al. [6] used neural networks for detecting air pollution exposure during walking or cycling trips. McCreddin et al. [7] developed an artificial neural network, Monte Carlo simulation, and other models to predict 24 h personal exposure to PM_{10}. In contrast to other papers, in this study neural networks were used independently to predict and classify the level of air pollution. Like Vakili et al. [8], we used multilayer perceptron with evaluation through the computation of root mean square error (RMSE). The knowledge that is obtained by applying neural network is the basis for creating a knowledge base (cf. [9]).

The Institute of Public Health in Čačak, Serbia has been monitoring the air quality in the area of the Moravica District for a number of years. Air quality control includes systematic monitoring of emission of basic and specific pollutants originating from stationary sources. It is conducted on a daily basis and includes the determination of daily concentrations of sulfur dioxide (SO_2), soot (black smoke) particles, nitrogen dioxide (NO_2) and total particulate matter. Soot, i.e., black smoke is produced by burning fossil fuels. These are fine, small particles of the size of about 5 μm which float in the air and behave like a gas. Soot contains toxic and carcinogenic substances and can accumulate bacteria. All these components easy penetrate and damage the respiratory system. Total particulate matter are pieces of solid fuel, ash and street dust which fall to the ground due to its weight. The effect on the organism depends on their origin and chemical composition, size and shape of the particles, contamination by microorganisms and heavy metals.

Air quality control is performed in order to determine:

- impact of sources of air pollution on the properties of air,
- degree of air load with pollutants,
- deviations of mechanisms of propagation of air pollution from its allowed level,
- insight into physicochemical and chemical processes influencing the transformation of primary pollutants,
- effect of the measures taken to prevent and/or reduce air pollution,

The aim of this study refers to determining the possibilities for application of artificial neural networks to predict the level of air pollution on the basis of input parameters. Moreover, users will be able to get insight into air pollution level based on new input data through the creation web-based application.

2. EXPERIMENTAL

Study area. The Moravica District is located in the western part of Central Serbia (Fig. 1). It consists of three municipalities: Lučani, Ivanjica and Gornji Milanovac and the town of Čačak (which is the administrative center of the Moravica District).

According to the 2011 census [10], Moravica District has the population of 212 149 and covers the area of 3016 km^2.

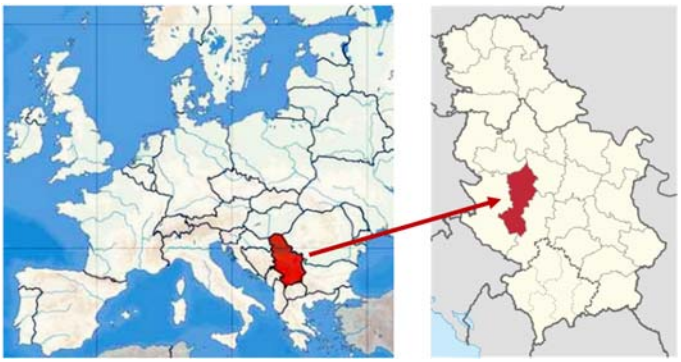

Fig. 1. Location of the Moravica District, Serbia ([11, 12])

The Moravica district has a developed agriculture, interesting tourist offer, high education and scientific institutions, favorable climate and environmental activities for the production of healthy food and good personnel potential. The climate of this part of Serbia is moderate continental. The average annual temperature is 10.47 °C. The coldest month is January with an average air temperature of –1.2 °C. The warmest month is July, with an average air temperature of 21 °C, so that the annual temperature amplitude is 22.2 °C. Central vegetation temperature is 16 °C, which is extremely favorable for the development of agriculture. The data used in this research was collected by the Institute of Public Health in Čačak, Serbia [17].

Sampling and measurements. The Institute of Public Health in Čačak, Serbia has continuously been measuring and analyzing the air pollution in Moravica District since 2005. Air sampling was conducted at the measuring points that are not directly exposed to the source of air pollution at the height of 1.5–10 m from ground level. The distribution of the measurement points depended on their location, layout and types of sources of pollution, population density, geography of terrain and weather conditions.Soot concentration was measured by the reflectometric method [13]. SO_2 and NO_2 concentrations were determined by spectrophotometric method [14, 15]. Particulate matter was measured by the method presented by Ramzin [16].

3. CREATION OF THE NEURAL NETWORK MODEL

The steps of creation of the artificial neural network model include: data selection, data preprocessing, modelling of neural network and testing the network through the data-mining extensions (DMX) queries.

3.1. DATA SELECTION AND PREPROCESSING

Table 1 shows the annual mean concentrations of air pollutants (SO_2, NO_2, soot and particulate matter) for ten measuring points in the Moravica District: 5 in Čačak, 3 in Ivanjica and 2 in Lučani. Limit and tolerable concentrations per year in Serbia (according to [18]), as well as the coordinates of each sampling site are also presented in the table. Various countries in Europe have different limit values for given pollutants, however for the soot particles, no limit concentrations have been established in the EU [19] so far.

Table 1

Mean annual concentrations of air pollutants at 10 locations

Location	Site	Coordinates		Soot [μg/m³]	SO_2 [μg/m³]	NO_2 [μg/m³]	Particulate matter [μg/(m³·day)]	Air pollution level
		[N]	[E]					
Čačak	1	43°53′32.0″	20°21′01.2″	24.38	25.73	48.78	121.82	2
Čačak	2	43°53′33.9″	20°22′01.1″	20.24	17.1	27.14	94.51	1
Čačak	3	43°53′32.0″	20°21′58.2″	32.12	24.75	27.1	267.63	2
Čačak	4	43°54′10.4″	20°20′40.7″	23.21	14.83	–	128.34	1
Čačak	5	43°53′38.9″	20°20′42.0″	34.05	33.42	–	87.39	1
Lučani	1	43°51′32.1″	20°08′09.7″	13.27	17.71	–	228.16	2
Lučani	2	43°51′40.9″	20°08′50.0″	–	–	–	192.14	1
Ivanjica	1	43°35′22.9″	20°13′32.6″	29.82	13.63	–	134.63	1
Ivanjica	2	43°34′31.6″	20°14′14.8″	35.05	19.67	–	101.93	1
Ivanjica	3	43°34′52.4″	20°13′47.8″	44.87	22.08	–	131.42	1
Limit value in Serbia [11]				50	50	40	200 μg/m³	
Tolerable value in Serbia [11]				75	50	60	–	

Air pollution level is defined according to the Law on Air Protection in Serbia [20]. According to the prescribed limit and tolerable values, the following air pollution categories have been determined:

1. I category – clean or slightly polluted air where limit values are not exceeded for any pollutant.
2. II category – moderately polluted air where limit values are exceeded for one or more pollutants, but the tolerable values are not exceeded for any pollutant.
3. III category – heavily polluted air where tolerable values for one or more pollutants are exceeded.

Data do not require special preprocessing and transformation because they are given in a suitable form for further analysis.

3.2 MODELLING, EVALUATION AND TESTING THE NEURAL NETWORK

A multilayer perceptron, which is one of the back-propagation neural networks, has been applied. It is a multilayer algorithm, and learning has been monitored. The back-propagation training algorithm has been used. The neural network algorithm is used to create a network that, in this study, can contain three layers of neurons: an input layer, a hidden layer (which is optional), and an output layer.

- Input layer (34.02.07 in ISO/IEC 2382-34:1999 [21]) defines all the input attribute values for the data mining model. Besides the soot, SO_2, NO_2 and particulate matter concentrations, the input values for the model are: year, municipality and measuring site.
- Hidden layer (34.02.10 in ISO/IEC 2382-34:1999, [21]) contains neurons which receive input from the input layer and forward them to the output layer. In this layer, weights are assigned to input neurons.
- Output layer (34.02.08 in ISO/IEC 2382-34:1999, [21]) contains neurons which represent the attribute values that we are predicting. In the approach used here, air pollution level is the output from the network. Figure 2 shows the structure of the neural network used.

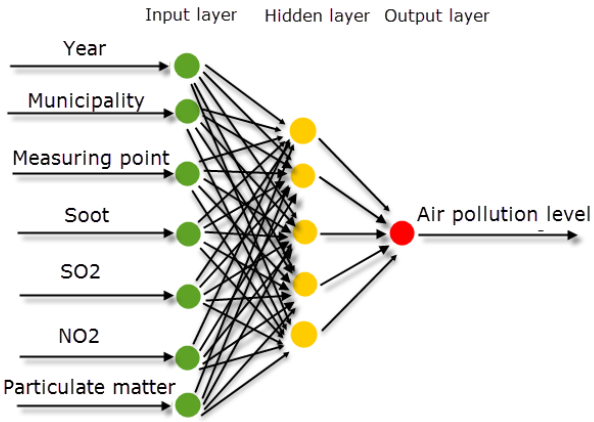

Fig. 2. Structure of the neural network

For the model evaluation, 30% of data for testing and 70% of data for training the neural network were used. The model was trained with the data obtained from the Institute of Public Health in Čačak, while it should be tested with new data that had not been used during training. New data could be collected in the coming years. Also, a lift chart was used to perform the evaluation here. A lift chart is a method of visualizing the improvement obtained by using the data from the analysis of the data model as compared to the randomly selected results.

For the evaluation purposes, the root mean square error (RMSE) has been used according to [22].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(t_i - o_i\right)^2} \tag{1}$$

where t_i is the calculated output that the network gives, o_i is the real output for the case i, and n is number of cases in the sample.

For the validation of the neural network model we also used cross validation by which the training and testing of the classifiers was performed. It is a model validation technique for assessing how the results will generalize to an independent data set. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

Testing the neural networks was conducted via data mining extensions (DMX) query [23]. The aim of using the DMX query was related to creating queries against the model in order to obtain the desired results, and an answer to the research questions was obtained. Here, the goal was to obtain the predicted probability of access to certain modules. DMX queries can be created within the Microsoft Visual Studio 2008 using a wizard, or queries could be written directly within the Microsoft SQL Server Management Studio 2008 after the selection of models and types of queries. We used singleton queries for the predictions.

4. RESULTS AND DISCUSSION

The value of the root mean square error was 0.0635. When the RMSE < 1, the model is useful. Lower RMSE values indicate more accurate models than the unintelligent predictor. The RMSE presents a relation between the total error of the model and its unintelligent predictor (which always predicts the mean value of the output). Figure 3 shows the lift chart for this study. The ideal line is a straight one, while the partly curved line indicates the actual accuracy of the predictions. Based on the chart, we can analyze the accuracy of the created models.

It can be concluded that 82% of the total population used to test the created data-mining model accurately predicted 72.73% of cases. Bearing in mind that the ideal model predicted 82% of correct cases out of the 82% of the population, the created model had a deviation of 9.27% from the ideal model. For this reason, the created model can be regarded a fairly accurate one. The value of 92% score is significant when two or more models are created and compared (which was not the case in this study).

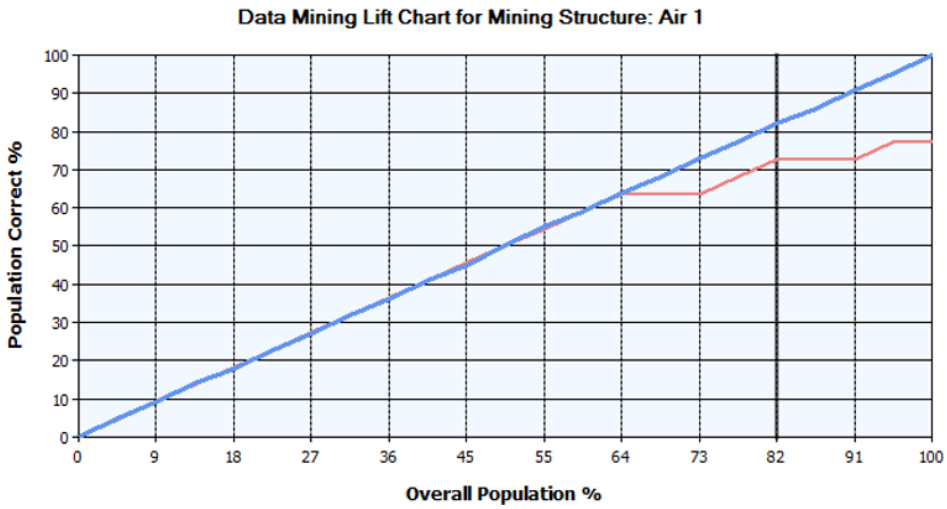

Fig. 3. Lift chart for the neural network model: ideal model – straight line, score 0.92, population correct 72.73%; air – line curved for the population higher than 64%, population correct 82.00%

The data set was divided at random into a set of K distinct sets. Training was performed on $K - 1$ set and the remaining set was tested. This was repeated for all of the possible K training and test sets. Average of all K results were the classification results. Testing of the neural network was conducted through the DMX singleton queries. The presented query gave the possibility of entering new input parameters, for which the neural network needed to assess the degree of contamination. Input data were examples of new data that would be measured in the coming years. The created neural network was tested through this approach.

In this case, for the measuring point 1 in Čačak for the year 2017 and for the given levels of measured parameters of air pollution, the obtained result was 2. This means that it was estimated that the contamination will be of moderate intensity. Predictions for the other measuring points could be made in a similar way. The first query uses the function Predict. In the above query new data, which was not available to the model during the phase of training, was entered. Values of 20.24, 34, 54,3 and 190 for soot, SO_2, NO_2 and particulate matter respectively were entered. Apart from the mentioned query, queries which use the PredictHistogram function were utilized. In this case, the input data were the measuring point 2 in the municipality of Lučani in 2017 with the given levels of air pollutants: 15 for soot, 46 for SO_2, 30.1 for NO_2 and 125 for particulate matter.

The pollution level obtained for the given input data and PredictHistogram function was 1 (no air pollution). A set of additional data was also obtained:

- support: the number of cases that support this result – 36,
- probability: probability that air pollution will be 1 for the given input data 97,4%.

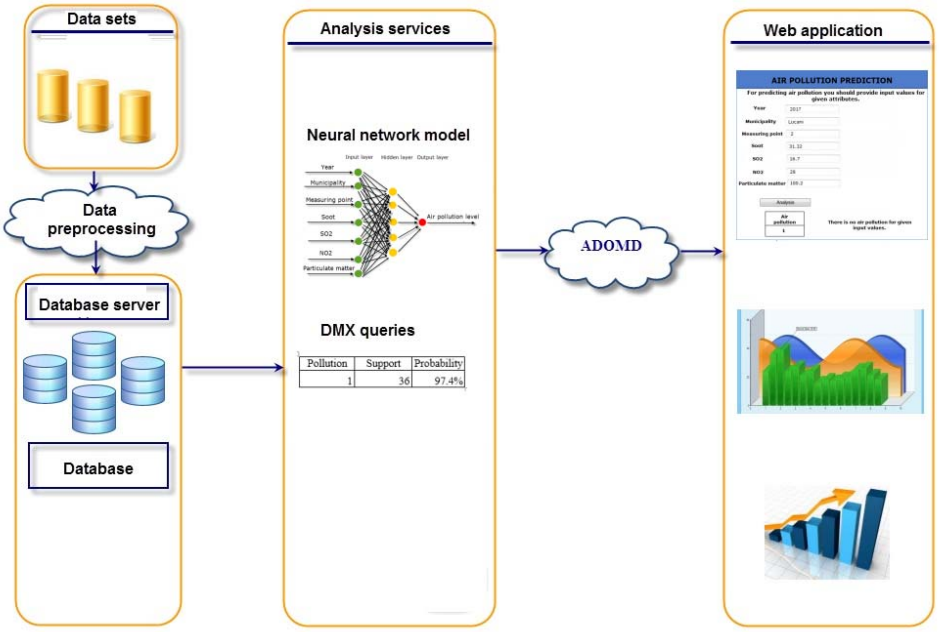

Fig. 4. System architecture [24]

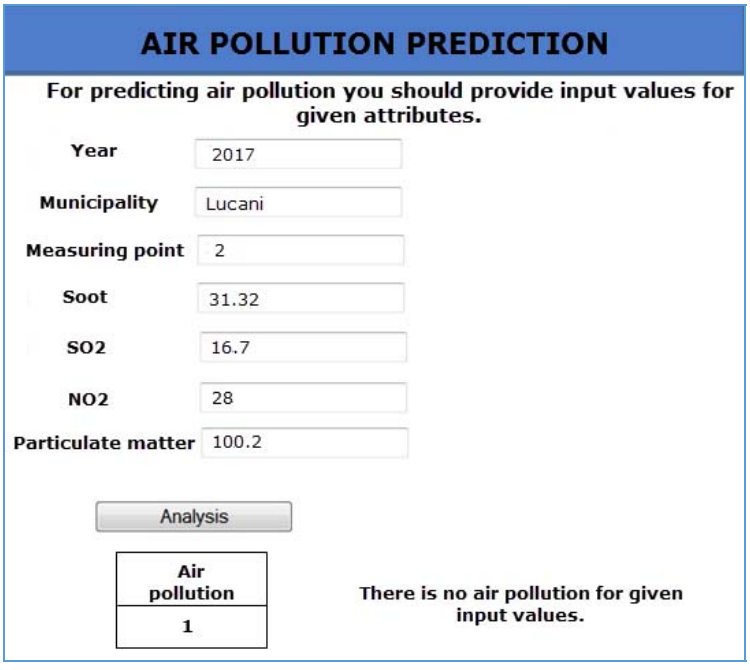

Fig. 5. Display of the user interface

The presented queries are only part of queries for created neural network model which show the options that customers can use within the tests. With the entry of new data neural network predicts the level of air pollution with satisfactory accuracy.

Besides the displayed query, the Web-based application was created for the target group of end users who may be persons with basic knowledge and skills in information technology. It was necessary to ensure that these end users can obtain their results through a simple form, without the knowledge of neural networks and DMX queries. The application was developed in the .NET programming environment, using the program language C#. The application was connected to the Microsoft SQL service for analysis through the ADOMD.NET. The ADOMD.NET is an environment of Microsoft.NET that enables communication with Microsoft SQL services for analysis. Figure 4 shows the application architecture which is similar to the one in the research by Blagojević and Micić [24]. Figure 5 shows the appearance of the user interface, and also one use case) with input variables and a result. The user enters the values for input attributes. By pressing the Analysis button, the results for air pollution are presented.

5. CONCLUSION

- The significance of the application of neural networks is reflected in the universality of access, relatively good accuracy of predictions and possible expansion of the model.
- The created neural network has satisfactory accuracy and the web-based application can be used for obtaining the level of air pollution for given input parameters.
- The advantage of the proposed approach is the possibility of testing the neural network with new data owned by users. A significant advantage of the proposed model is that it can be expanded to include more air pollutants and other input parameters and that the application, in addition to being simple to use and able to be used by an average PC user, enables the integration of new DMX queries.
- Shortcomings of the application, related to tutorials and the improvement of visualization of results, will be removed during the future work. In addition, inclusion of other data on mining techniques is planned to predict air pollution .

ACKNOWLEDGMENT

This study was supported by the Serbian Ministry of Education and Science, Project III 44006.[1]

[1]http://www.mi.sanu.ac.rs/projects/projects.htm#interdisciplinary

REFERENCES

[1] PATRONAS D., KARIDA A., PAPADOPOULOU A., PISIHA A., XIPOLITOS K., KOKKINIS G., VOSNAIKOS K., GRAMMATIKIS B., VOSNIAKOS F., VASDEKIS K., *Air pollution and noise pollution due to traffic in three Greek cities*, J. Environ. Prot. Ecol., 2009, 10 (2), 332.
[2] HAIDUC C., ROBA L., BOBOS L., FECHETE-RADU L., *Urban aerosols pollution. Case study: Cluj–Napoca City*, J. Environ. Prot. Ecol., 2009, 10 (3), 611.
[3] HE G., FAN M., ZHOU M., *The effect of air pollution on mortality in China: Evidence from the 2008 Beijing Olympic Games*, J. Environ. Econ. Manage., 2016, 79, 18.
[4] XIAO S., LIU R., WEI Y., FENG L., LVA X., TANG F., *Air pollution and blood lipid markers level. Estimating short- and long-term effects on elderly hypertension inpatients complicated with or without type 2 diabetes*, Environ. Pollut., 2016, 215, 135.
[5] POPESCU F., IONEL I., LONTIS N., CALIN L., DUNGAN I.L., *Air quality monitoring in an urban agglomeration*, Rom. J. Phys., 2011, 56, 495.
[6] ADAMS M., YIANNAKOULIAS N., KANAROGLOU P., *Air pollution exposure: An activity pattern approach for active transportation*, Atmos. Environ., 2016, 140, 52.
[7] MCCREDDIN A., ALAM M.S., MCNABOLA A., *Modelling personal exposure to particulate air pollution: An assessment of time-integrated activity modelling, Monte Carlo simulation and artificial neural network approaches*, Int. J. Hyg. Environ. Health, 2015, 218 (1), 107.
[8] VAKILI M., SABBAGH-YAZDI S., KALHOR K., KHOSROJERDI S., *Using Artificial neural networks for prediction of global solar radiation in tehran considering particulate matter air pollution*, Energy Proc., 2015, 74, 1205.
[9] CHAN K., JIAN L., *Identification of significant factors for air pollution levels using a neural network based knowledge discovery system*, Neurocomp., 2013, 99, 564.
[10] *2011 Census of Population, Households and Dwellings in The Republic of Serbia*, Comparative overview of the number of population in 1948, 1953, 1691, 1971, 1981, 1991, 2002 and 2011, Statistical Office of the Republic of Serbia, Belgrade, 2014.
[11] *The neoliberal paradigm and winds from Olympus* http://www.nspm.rs/images/stories/01maja/aa2/Srbija-u-Evropi.jpg, retrieved on 26/9/2016
[12] *Moravica district*, https://sr.wikipedia.org/wiki/Моравички_управни_округ#/media/File:Moraica_in_Serbia.svg, retrieved 26/9/2016.
[13] ISO 9835:1993, *Ambient air – Determination of a black smoke index.*
[14] SRPS ISO 6767:1997. *Ambient air – Determination of the mass concentration of sulfur dioxide – Tetrachloromercurate (TCM)/pararosaniline method*.
[15] *Nitric oxide and nitrogen dioxide. Method 6014*, *Issue 1*, 4th Ed., NIOSH Manual of Analytical Methods (NMAM), 1994.
[16] RAMZIN S., *Particulate matter (air). Determination of soluble, insoluble matter and ash VMK 043. Manual for communal hygiene*, Medicinska Knjiga, Beograd 1966.
[17] *Statistics of the Institute of Public Health in Čačak*, *Serbia*, http://www.zdravljecacak.org/stranice/Statisticki%20god.php, retrieved 19/7/2016.
[18] *Book of regulations on conditions and requirements for monitoring of air quality*, Official Gazette of the Republic of Serbia, No. 11/2010.
[19] *Clean Air Copenhagen. Air Quality Challenges and Solutions*, The Danish Ecological Council, 2014.
[20] *Law on Air Protection in Serbia*, Official Gazette of the Republic of Serbia, Nos. 36/2009 and 10/2013
[21] ISO/IEC 2382-34:1999 *Information technology – Vocabulary – Part 34: Artificial intelligence – Neural Networks*, 1999.

[22] DRAPER C., REICHLE R., JEU R., NAEMI V., PARINUSSA R., WAGNER W., *Estimating root mean square errors in remotely sensed soil moisture over continental scale domains*, Remote Sens. Environ., 2013, 137, 288.
[23] *Data mining extensions queries*, https://msdn.microsoft.com/en-us/library/ms174788.aspx, retrieved 26/7/2016
[24] BLAGOJEVIC M., MICIĆ Ž., *Web-based intelligent report e-learning system using data mining techniques*, Comp. Electr. Eng. 2013, 39 (2), 465.