Science of the Total Environment 443 (2013) 511–519

Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization

Davor Z. Antanasijević [a,*], Viktor V. Pocajt [b], Dragan S. Povrenović [b], Mirjana Đ. Ristić [b], Aleksandra A. Perić-Grujić [b]

[a] University of Belgrade, Innovation Center of the Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia
[b] University of Belgrade, Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia

HIGHLIGHTS

- Neural network (ANN) modeling of annual PM_{10} emissions at a national level
- Sustainability and economical/industrial parameters are used as model inputs.
- The selection of inputs was based on smoothing factor (ISF) calculated by GA.
- The ANN model provides much better results in comparison with conventional models.
- Up to two years forecast with the ANN model can be made successfully and accurately.

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form 18 October 2012
Accepted 25 October 2012
Available online 4 December 2012

Keywords:
Neural networks
Multiple linear regression
Principal component regression
Annual PM_{10} emission forecasting

ABSTRACT

This paper describes the development of an artificial neural network (ANN) model for the forecasting of annual PM_{10} emissions at the national level, using widely available sustainability and economical/industrial parameters as inputs. The inputs for the model were selected and optimized using a genetic algorithm and the ANN was trained using the following variables: gross domestic product, gross inland energy consumption, incineration of wood, motorization rate, production of paper and paperboard, sawn wood production, production of refined copper, production of aluminum, production of pig iron and production of crude steel. The wide availability of the input parameters used in this model can overcome a lack of data and basic environmental indicators in many countries, which can prevent or seriously impede PM emission forecasting. The model was trained and validated with the data for 26 EU countries for the period from 1999 to 2006. PM_{10} emission data, collected through the Convention on Long-range Transboundary Air Pollution — CLRTAP and the EMEP Programme or as emission estimations by the Regional Air Pollution Information and Simulation (RAINS) model, were obtained from Eurostat.
The ANN model has shown very good performance and demonstrated that the forecast of PM_{10} emission up to two years can be made successfully and accurately. The mean absolute error for two-year PM_{10} emission prediction was only 10%, which is more than three times better than the predictions obtained from the conventional multi-linear regression and principal component regression models that were trained and tested using the same datasets and input variables.

1. Introduction

Nearly all governments have committed themselves to sustainable development by integrating economic welfare, environmental quality and social coherence (Böhringer and Jochem, 2007). Sustainable development provides an effective response to the socio-economic needs and interests of people and at the same time it eliminates or significantly reduces risks and threats to the environment and natural resources. However, many countries lack data necessary to use models for the assessment and forecasting of sustainable development indicators, environmental impact and pollutant emissions, which are important for environmental management and implementation of strategies for improving the environment. It has been shown that, in the absence of such initial data, widely available sustainability and economic performance indicators can be used as an input for artificial neural networks (ANNs), for prediction of various environmental indicators at the national level (Sözen et al., 2009; Antanasijević et al., in press; Radojević et al., in press). These indications can then be used to support the decision making process on sustainable development and environmental quality upgrade at the regional, national and international levels.

Emission data is one of the principal driving forces in air quality modeling (Dennis et al., 2010; Ciancarella et al., 2011). Also, many studies have shown that emission input to air quality models is one of the main

* Corresponding author. Tel.: +381 11 3303 642; fax: +381 11 3370 387.
E-mail address: dantanasijevic@tmf.bg.ac.rs (D.Z. Antanasijević).

0048-9697/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2012.10.110

sources of uncertainty (Russell and Dennis, 2000; Hanna et al., 2001). The main reason of emission data uncertainty is the method of their calculation (the multiplication of activity data with emission factors), because emission factors, depending on the determination method, have the minimum uncertainty of 10 to 30% (EEA, 2009).

ANN models have been used recently in many studies for forecasting air pollutant emissions. Kassomenos et al. (2006) described development of an ANN model for the calculation of road emissions for PM_{10} and four other pollutants. Lim et al. (2007) used ANN model for prediction of ammonia emission from field-applied manure. Karacan (2008) applied ANNs to predict the methane emissions from ventilation of U.S. longwall mines. Sözen et al. (2009) and Radojević et al. (in press) used ANN models to forecast the GHG emissions in Turkey and Serbia, respectively.

Particulate matter (PM), which is primarily generated by power generation, industry and transportation, stands out as one of the most important pollutants, since it induces serious health consequences for the exposed population all over the world.

Many epidemiological and panel studies have shown an association between the levels of particulate matter in urban air and short-term cardiopulmonary effect (Scapellato et al., 2009). PM is regarded as a top priority pollutant by the European Framework Directive that primarily seeks to protect human health from air pollutants by improving the air quality (Sfetsos and Vlachogiannis, 2010).

PM_{10} emission predictions are relevant to existing legislation in the EU, namely the 2001/81/EC — NEC directive (EC, 2001) and also to the protocols of the Geneva Convention (Long Range Transboundary Air Pollution—LRTAP), which most European countries have signed and ratified. Reporting obligations for both EU legislation and Geneva Convention include projections on future emissions of air pollutants. Many countries are already reporting PM_{10} emissions and projections under UNECE/LRTAP. PM_{10} is not included in the NEC Directive, but the methodology can be applied to the other pollutants, as well.

This paper describes the development of an ANN model for the prediction of annual PM_{10} emissions at the national level, using widely available sustainability and economical/industrial parameters as input. Original methods for improving the results of modeling will be described as well, including the application of genetic algorithm for optimizing the process of selection of input variables. In the evaluation section, the results of the ANN models are compared with conventional multiple linear regression (MLR) and principal component regression (PCR) models, using multiple performance criteria.

2. A comparison of the emission inventory based approaches and the ANN based approach to PM_{10} emission modeling

A usual approach to PM_{10} emission forecasting is based on the use of emission inventories. The bulk of an emission inventory is compiled by collecting activity data and appropriate emission factors according to the Tier 1 default approach:

$$\text{Emission}_{\text{pollutant}} = \sum_{\text{activities}} \text{Activity} \times \text{rate}_{\text{activity}} \times \text{Emission} \times \text{factor}_{\text{activity,pollutant}}. \quad (1)$$

Activity data is usually derived from economy statistics, including energy statistics and balances, economic production rates, population data, etc. (EEA, 2009). Emission factors must be determined for every sector, separately, either by measurement or calculation. Emission factors, determined by estimation based on a large number of measurements made at a large number of facilities that fully represent the sector, have uncertainty ranges of 10 to 30% (EEA, 2009).

The main emission factor guidebooks are:

EMEP/EEA guidebook (formerly referred to as the EMEP CORINAIR) which provides guidance on estimating emissions from both anthropogenic and natural emission sources. It is designed to facilitate reporting of emission inventories by countries to the UNECE Convention on Long-range Transboundary Air Pollution (CLRTAP) and the EU National Emission Ceilings Directive (EEA, 2009);

AP42 (published since 1972) is the primary compilation of EPA's emission factor. It contains emission factors and process information for more than 200 air pollution source categories (US EPA, 1995).

Eurostat obtained the data on air pollution from the reports under CLRTAP and under the EMEP Programme (Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air pollutants in Europe). Where PM_{10} data are not reported by countries to EMEP/CLRTAP, emission estimates were obtained from the Regional Air Pollution Information and Simulation (RAINS) model (Eurostat, 2012). A detailed description of the RAINS model can be found in Schöpp et al. (1999) and GAINS EUROPE (2012a).

The RAINS model was used also for the calculation of particulate matter emission by the European Environment Agency (EEA). Since fine particles are emitted from a large number of highly diversified sources, the RAINS methodology distinguishes 392 source categories for stationary energy combustion, industrial processes, mobile sources and agriculture. In total, approximately 800 parameters are needed for emission forecasting (EEA, 2012).

Schöpp et al. (2005) conducted uncertainty analysis of emission estimates in the RAINS model. They concluded that the uncertainty of model emission predictions varies depending on country. Several factors contribute to these uncertainties: if emissions are dominated by one or two source sectors, there is only a limited potential that errors in the sector estimates can be statistically compensated by other; if parameters (activity data or emission factors) of a large sector are especially uncertain, this sector will influence strongly the uncertainty of national total emissions; for some countries it was assumed that the data are more uncertain than for the other, e.g., data for Central and Eastern European countries are more uncertain than for the EU15 (Schöpp et al., 2005).

Since year 2007 the RAINS model has been incorporated in the Greenhouse Gas and Air Pollution Interactions and Synergies (GAINS) model (GAINS EUROPE, 2012b).

The main difference between conventional (e.g. RAINS model) and the ANN approach is that the ANN model requires substantially smaller number of input parameters. Thus, the ANN model can be applied for PM_{10} emission forecasting when countries and regions cannot adequately predict input parameters needed for models based on activity levels and emission factors.

The proposed ANN model uses some of the performance indicators that have been used as activity data in emission inventory models, such as statistics of population, economical performance and energy consumption; however it significantly reduces the number of required input parameters compared to the emission inventory approach. Also, the proposed ANN model eliminates the need to determine emission factors, thereby reducing the uncertainty of input parameters.

3. Modeling techniques

3.1. Artificial neural networks

Artificial neural network (ANN) is a modeling technique which can determine non-linear relationships between variables in input datasets and variables in output datasets. ANNs modeling is based on a learning (training; calibration) process, after which the ANN network can estimate values of output variables for input datasets. ANNs need a considerable amount of historical data to be trained; upon satisfactory training, an ANN should be able to provide output for previously "unseen" inputs (Palani et al., 2008). Often, there can be some uncertainty about precisely which input variables to use. The selection of input variables for an ANN forecasting model is a key issue, since irrelevant or noisy variables may have negative effects on the training process, resulting to an

unnecessarily complex model structure and poor generalization power (Voukantsis et al., 2011).

Genetic algorithms (GAs) are frequently used in approximation problems for the selection of input variables and therefore their application reduces the total number of predictors (Grivas and Chaloulakou, 2006). The GA is a model of machine learning, which derives its behavior from a representation of the processes of evolution in nature. Detailed structure of a standard genetic algorithm can be found elsewhere (Kalogirou, 2003).

The ANN architecture used in this study is General Regression Neural Network (GRNN) (Specht, 1991; Kim and Kim, 2008; Palani et al., 2008), since a relatively small input dataset was available. GRNN is a three layer ANN consisting of a hidden layer that can be divided into two sub layers: the pattern layer and the summation layer (Fig. 1). The number of neurons in the input and output layer is equal to the number of input and output parameters that have been used in the ANN model. The number of neurons in the pattern sub layer is equal to the number of training cases used for model training.

The summation sub layer has two different types of processing units, the summation units and a single division unit. The number of the summation units is always the same as the number of the GRNN output units. The division unit only sums the weighted activations of the pattern units of the hidden layer, without using any activation function. Each of the GRNN output units is connected only to its corresponding summation unit and to the division unit (there are no weights in these connections) (Kalogirou, 2003).

The GRNN network uses supervised training and compares its resulting outputs against measured outputs, which were presented to the network in the training data set. The resulting parameter of the GRNN training is a smoothing factor, which is applied when the GRNN network is applied on a specific dataset. Smoothing factor represents the width of the calculated Gaussian curve for each probability density function. The GRNN model trained with a genetic algorithm (GA) gives the individual smoothing factors (ISFs) for each input, as well as an overall smoothing factor. When GA is used, overall smoothing factor value is calculated by multiplying individual smoothing factor (ISF) value of each input. The ISFs can be used as a sensitivity analysis tool: the larger the factor for a given input is, the more important the input is to the model.

3.2. Conventional models

The multiple linear regression (MLR) method is a commonly used technique to obtain a linear input–output model for a given data set. The multiple regression approach can face serious difficulties when the independent variables are correlated with each other. Multicollinearity, or high correlation between the independent variables in a regression equation, can make it difficult to correctly identify the most important contributors to a physical process (Abdul-Wahab et al., 2005). Multicollinearity can be detected by computing correlations between all pairs of inputs or calculating the variance inflation factors (VIF) for each model input. One method of removing such multicollinearity from the independent variables is to use the principal component analysis (PCA) (Al-Alawi et al., 2008).

Principal Components Regression (PCR) is a combination of PCA and MLR. The PCR can be divided into three consecutive steps: first a PCA is run on the table of input variables, then a MLR is run on the components selected by PCA in the first step, and lastly the parameters of the model, that correspond to the input variables, are computed.

4. Methodology

4.1. Input and output variables and data

During the process of definition of variables, the availability of input and output datasets must be considered. Wide availability of sustainability and economical/industrial parameters makes them suitable for the creation of environmental ANN models; several studies have already shown that such ANN models can achieve good accuracy (Sözen et al., 2009; Antanasijević et al., in press; Radojević et al., in press).

Gross domestic product (GDP), gross inland energy consumption (GIEC) and motorization rate were selected as the basic input variables for the PM_{10} emission model. The GIEC and motorization rate were chosen since they represent power generation and transportation, which are two of the three main PM_{10} emission sources, GDP shows the size of economy, therefore indicating the size of industry, which is the third main source of PM_{10} emission. Also, available data for diversified industry sectors, which are known as PM_{10} emission sectors, are chosen as inputs. The variables chosen for inputs and output of models, with their units after normalization and data sources, are presented in Table 1. The relationship between the chosen model input variables and PM_{10} emission is presented in Fig. 2.

The same PM_{10} emission data (for years 1999, 2004 and 2006) with mean values and the change of emission for the period 2004–2006 and for the studied period are presented in Table 2. To create training, validation and test datasets, available data for the European Union countries, excluding Malta (because of the lack of data), and the European Union as a whole (EU27) was used.

These datasets are defined as follows:

- training dataset — the group of data used for GRNN model training; in case of MLR and PCR models represents the group of data used for model creation;
- validation dataset — the group of data, extracted from the training

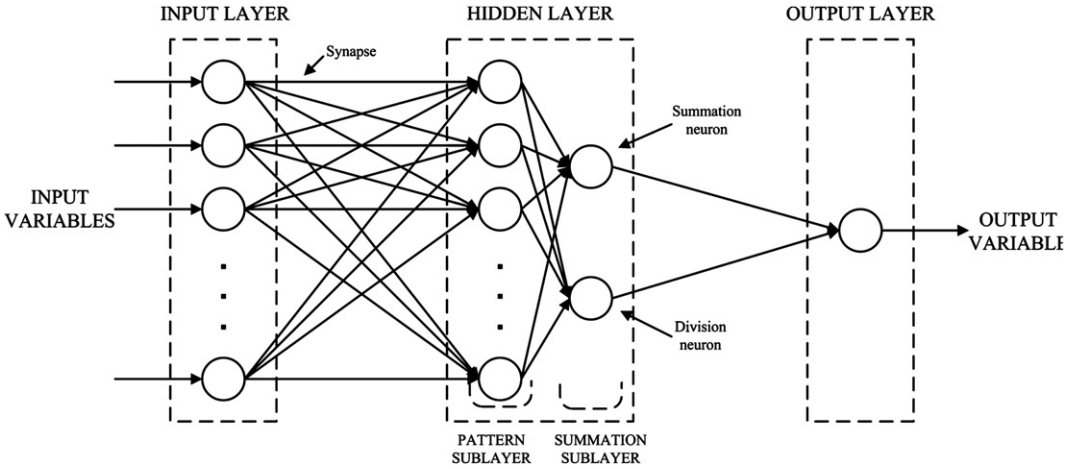

Fig. 1. General regression neural network architecture.

Table 1
List of available variables.

Label	Input variable	Unit[a]	Data source
V1	Gross domestic product (GDP)	(–)[b]	Eurostat (Eurostat, 2011)
V2	Gross inland energy consumption	toe pc[c]	
V3	Incineration of wood	m^3 pc	
V4	Motorization rate	cars pc	
V5	Primary production of coal and lignite	toe pc	
V6	Production of paper and paperboard	t pc	
V7	Roundwood production	m^3 pc	
V8	Sawnwood production	m^3 pc	
V9	Production of refined copper	t pc	U.S. Geological Survey (USGS, 2011)
V10	Production of aluminum	t pc	
V11	Production of pig iron	t pc	International Iron and Steel Institute (IISI, 2011)
V12	Production of crude steel	t pc	
V13	Production of NPK fertilizers	t pc	International Fertilizer Industry Association (IFA, 2011)
	Output variable		
PM_{10}	PM_{10} emission	t pc	Eurostat (Eurostat, 2011)

[a] After normalization.
[b] Unitless because of normalization per GDP value of EU27.
[c] The tons of oil equivalent (toe) per capita.

dataset and used in training process to prevent ANN overtraining and to enable better generalization of the ANN model on new data (also used for the model validation during PCR model creation);
- test dataset — the group of new data presented to the ANN model after the training process, used to evaluate ANN model generalization (same as in case of MLR and PCR model development).

Since ANNs generally achieve better performance with normalized values, all selected variables were normalized per capita. GDP was also normalized per GDP value of EU27 as a group. Since the PM_{10} emission data on Eurostat (2011) was published from the year 1999 to 2006, the data from 1999 to 2004 was used as the training and validation datasets. The data from the years 2005 and 2006 was used as the test dataset.

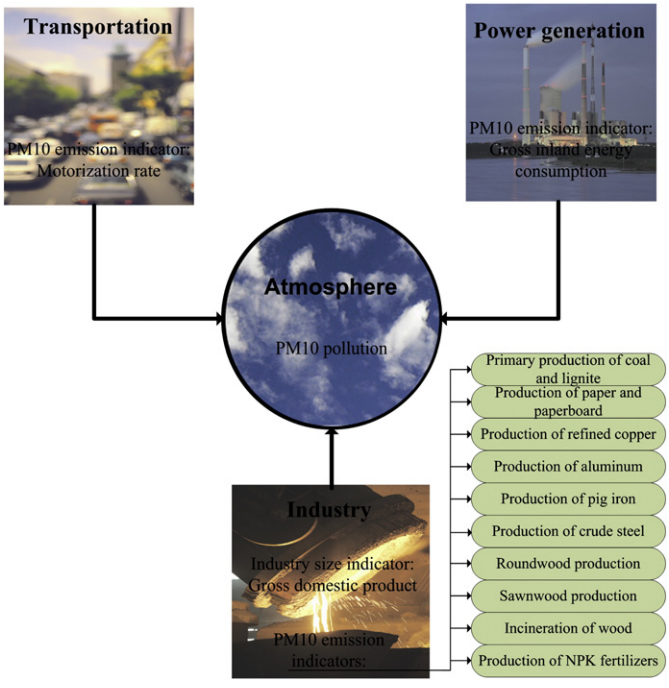

Fig. 2. Relationship between the chosen model input variables and PM_{10} emission.

Table 2
PM_{10} emission (mean, 1999, 2004 and 2006 values) and the change of PM_{10} emission for study period and for period 2004–2006.

Country/region	PM_{10} emission[a] (kg pc)				Change of emission (%)	
	Mean	Year			Period	
		1999	2004	2006	1999–2006	2004–2006
EU 27	42.44	47.53	40.53	38.48	−19.0	−5.05
Austria	36.88	35.25	37.59	36.23	2.79	−3.60
Belgium	43.89	44.43	42.43	38.64	−13.0	−8.93
Bulgaria	104.1	98.97	105.7	106.9	7.99	1.07
Cyprus	66.27	75.01	61.70	50.91	−32.1	−17.5
Czech Republic	50.08	57.08	49.23	42.69	−25.2	−13.3
Denmark	55.40	61.70	N/A	50.18	−18.9	N/A
Estonia	85.61	90.73	86.23	67.61	−25.5	−21.6
Finland	58.18	61.37	58.09	55.85	−8.98	−3.85
France	42.68	48.52	40.77	37.87	−22.0	−7.11
Germany	28.88	33.20	27.42	25.72	−22.6	−6.21
Greece	63.11	64.74	N/A	N/A	N/A	N/A
Hungary	43.42	57.39	38.71	34.38	−40.1	−11.2
Ireland	66.07	79.17	56.81	N/A	N/A	N/A
Italy	33.05	39.81	30.36	26.83	−32.6	−11.6
Latvia	26.42	29.60	26.08	28.29	−4.41	8.47
Lithuania	30.15	32.05	29.93	32.52	1.45	8.63
Luxembourg	23.60	29.02	21.39	14.90	−48.6	−30.3
Netherlands	29.84	34.39	28.11	26.35	−23.4	−6.25
Poland	51.61	59.13	48.71	49.70	−16.0	2.04
Portugal	59.66	61.47	N/A	N/A	N/A	N/A
Romania	33.33	29.76	N/A	42.89	44.1	N/A
Slovakia	50.86	64.95	35.67	33.52	−45.8	−6.02
Slovenia	38.93	48.21	45.22	35.23	−30.5	−22.1
Spain	60.16	64.26	59.19	54.44	−15.3	−8.02
Sweden	31.82	36.56	30.16	28.32	−22.5	−6.09
United Kingdom	40.98	48.18	37.83	35.12	−27.1	−7.16

N/A — missing value or not applicable.
[a] Eurostat (2011).

4.2. GRNN model development

Both the development and evaluation of the model were carried out in accordance with "good model-development practice" proposed by Jakeman et al. (2006). The first step in GRNN model development is to divide training dataset into two datasets: actual training dataset and validation dataset (used for internal network validation during the training process in order to prevent network overtraining). Validation dataset is extracted randomly choosing 24 data patterns (16%) from available 154.

The GRNN architecture parameters (number of neurons in input, output and hidden layer) are set automatically, since they depend on the number of input and output variables and number of training data patterns (as described in section 3.2.), and so these parameters are defined by the training dataset presented to the network. The only parameter that can be modified at this stage is scale function. All tested GRNN models had 1 neuron in the output layer, 154 neurons in the hidden layer (that is equal to the number of training data patterns) and the linear scale function, while the number of neurons in the input layer varies from 7 to 13.

The Euclidian distance metric method, which is recommended for most networks for computing the distance between patterns during network training, was selected, since the GRNN network works by comparing patterns based upon their distance from each other. Another important GRNN training parameter is the algorithm for finding the best smoothing factor, which defines GRNN's predicting performance. The determination of smoothing factor can be done manually,

or by using an iterative or genetic algorithm. In the present study, genetic algorithm (GA), which is recommended when the input variables don't have the same impact on predicting the output, was used.

GA uses a "fitness" measure to determine which of the individuals in the population survive and reproduce. Thus, survival of the fittest causes good solutions to evolve. The fitness measure for GRNN is the mean squared error of the outputs over the entire validation set. The whole range of ISF is tested by GA until the combination of ISF that works the best on the validation dataset is found.

Input variable selection was based on the ISF which is calculated through the GRNN training by GA for each input variable. The ISF value ranges between 0 (non-significance) and 3 (the highest significance) which enables the elimination of less significant variables. The ISF input variable selection method was divided in two consequent procedures: the elimination procedure and add-on procedure. Schematic representation of the elimination procedure of input variable selection is presented in Fig. 3.

Although ISF values were changed whenever any of the input variables had been eliminated and a new GRNN model was created, this is an efficient way to eliminate non-significant input variables. It is important to emphasize that the criterion for the termination of the optimization procedure was the performance comparison of the last created GRNN model with the non-optimized GRNN model, trained with all available input variables. The add-on procedure is based on the values of ISF, determined during the elimination procedure. The variables that had an ISF larger than 2 in any of the created GRNN models (Table 3) were selected as "base" variables. The variables that had an ISF value from 1 to 2 in any of the created GRNN models were selected as "additional" variables (Table 3). One of the GRNN models created by this procedure had only "base" variables as inputs, while the other four GRNN models were created using the "base" variables and different combinations of "additional" variables.

5. Results and discussion

5.1. Performance evaluation

The analysis of the GRNN model's performance is based on a statistical comparison of the GRNN model results with the actual PM_{10} emissions and with the results of other models. The ability of the models to produce accurate results was measured using the following statistical metrics:

- The root mean squared error (*RMSE*)

$$RMSE = \left[\overline{\left(C_p - C_o\right)^2}\right]^{1/2} \quad (2)$$

- The normalized mean squared error (*NMSE*)

$$NMSE = \frac{\overline{(C_O - C_P)}}{C_O C_P} \quad (3)$$

- The mean absolute error (*MAE*)

$$MAE = \frac{1}{n}\sum\left|C_p - C_o\right| \quad (4)$$

- The correlation coefficient (r)

$$r = \frac{\overline{\left(C_O - \overline{C_O}\right)\left(C_P - \overline{C_P}\right)}}{\sigma_{C_O}\sigma_{C_P}} \quad (5)$$

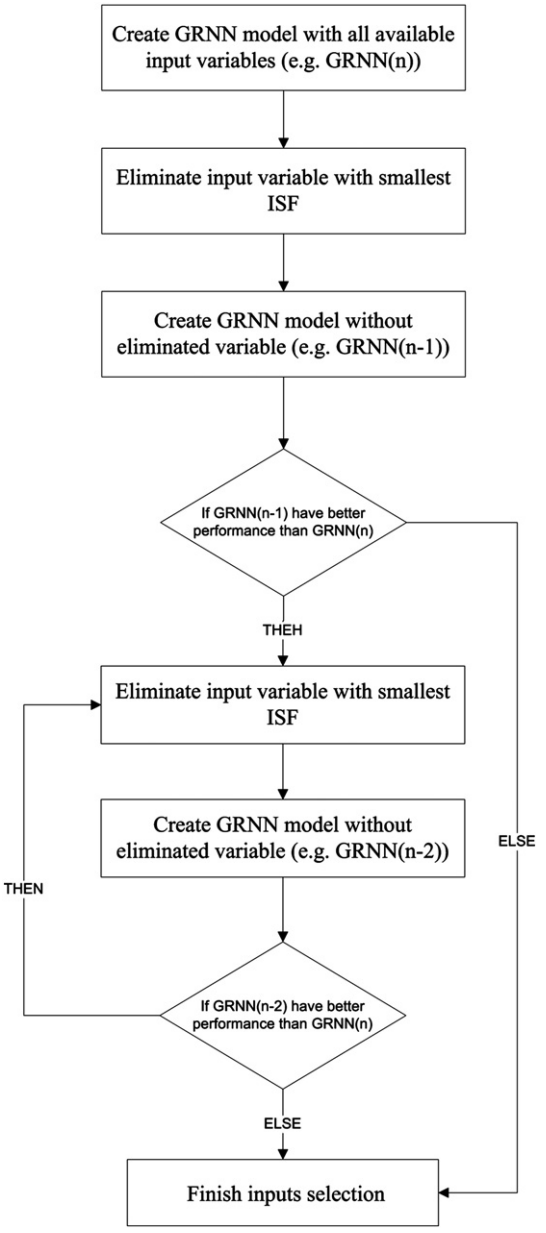

Fig. 3. Schematic representation of the elimination procedure of ISF input variables selection.

- The index of agreement (*IA*)

$$IA = 1 - \frac{\overline{\left(C_p - C_o\right)^2}}{\overline{\left[\left|C_p - \overline{C_o}\right| + \left|C_o - \overline{C_o}\right|\right]^2}} \quad (6)$$

- The fractional bias (*FB*)

$$FB = \frac{\overline{C_O} - \overline{C_P}}{0.5\left(\overline{C_O} + \overline{C_P}\right)} \quad (7)$$

where C_P and C_O are the predicted and observed concentrations, respectively, and σ_{C_O} and σ_{C_P} are the standard deviation of observations and predictions, respectively. The some of selected statistical metrics were previously used in the studies of PM_{10} and other air pollutant emission modeling by ANN and other modeling techniques (Najafi et al., 2009; Pay et al., 2012; Koo et al., 2012).

Table 3
The values of ISFs (sensitivity) for input variables.

GRNN model	V1[a]	V2[b]	V3[b]	V4[a]	V5[c]	V6[a]	V7[c]	V8[a]	V9[a]	V10[a]	V11[b]	V12[a]	V13[c]
GRNN13	2.79	1.61	1.92	2.56	0,01	0.33	0.26	2.78	2.44	1.35	1.75	0.55	0.24
GRNN12	2.18	1.09	1.64	2.40	–	0.42	0.32	2.95	2.34	0.41	1.72	0.54	0.60
GRNN11	2.32	0.41	0.05	0.42	–	2.36	–	2.93	2.04	2.71	1.07	2.15	0.02
GRNN10	2.55	1.66	1.13	1.95	–	0.07	–	3.00	2.88	1.19	0.02	1.27	–
GRNN9	1.18	0.48	0.34	0.81	–	0.48	–	2.92	2.76	0.75	–	1.53	–

[a] "Base" variables.
[b] "Additional" variables.
[c] Ignored variables.

In addition to the listed statistical performance metrics, in this study was also used:

- The fractional variance (*FS*)

$$FS = \frac{\sigma_{C_O} - \sigma_{C_P}}{0.5\left(\sigma_{C_O} + \sigma_{C_P}\right)} \quad (8)$$

- The percent of predictions within a factor of 1.25 of the observed values (*FA1.25*)

$$0.8 < \frac{C_p}{C_o} < 1.25. \quad (9)$$

The *FS* and *FA* indicators were previously used for evaluation of air dispersion models for mercury emissions (Patel and Kumar, 1998).

It is important to emphasize that the significance of these statistical metrics is not equal. In principle, the most critical metric is *IA*, which is the measure of the correlation of the predicted and actual emission values and also incorporates the error between those values. In order of importance, *IA* is followed by: *FA1.25*, which represents the proportion of data for which the model results are "approximate" with the measured values; and then *FB*, which is a measure of the agreement of the mean emission values. The key statistical parameters for the models evaluation were therefore *IA* and *FA1.25*.

For the evaluation of model performance for each particular country, the mean absolute percentage error (*MAPE*) was used:

$$MAPE = 100\frac{1}{k}\sum\frac{\left|C_o - C_p\right|}{C_o} \quad (10)$$

5.2. Results of the GRNN model

The GRNN models and the values of ISF determined during the Elimination procedure are presented in Table 3.

A comparison of the created GRNN models and the conventional one is presented in Table 4.

The best predictions are achieved with the GRNN10 model (model with 10 inputs, the list of inputs is presented in Table 3.), which has an *IA* that ranges from 0.99 to 0.98 and *FA1.25* that ranges from 100% to 91%, for one year (2005) and two year (2005–2006) forecasts, respectively. Since the non-optimized GRNN model (GRNN13) has an *IA* that ranges from 0.99 to 0.96 and *FA1.25* that ranges from 100% to 89%, for one-year and two-year forecasts, respectively, the improvement of the GRNN model with variables optimization was achieved. Also, the optimized model has three inputs (23%) less than non-optimized model, which significantly reduces data preparation and also reduces the uncertainties when the model is used for future emission predictions with forecasted input variables.

A slight decrease of model accuracy for the year 2006 is a consequence of the changes related to the PM_{10} emission from the training period (1999–2004) compared to the forecasted year. For EU27, the decrease of PM_{10} emission during the studied period is 19%, while for some countries decrease is even higher than 30% and goes up to almost 50% (Table 2.). This PM_{10} emission decrease is related to the enforcement of the EU environmental directives, 96/61/EC (EC, 1996), 98/69/EC (EC, 1998), 99/30/EC (EC, 1999), and 01/80/EC (EC, 2001), which limited PM_{10} emissions during the studied period.

Based on sensitivity analysis (Table 3), the input variables of the GRNN10 model can be classified by relative importance in several groups: GDP and some of industrial emission indicators (sandwood and refined copper production) are the most important; in the second group are power generation and transportation indicators; and in the third group, as less important, are remaining industrial emission indicators.

Table 4
Performance indicators of created GRNN and conventional models.

Model	Forecast period	Statistical metrics for test dataset							
		IA	FA1.25	FB	NMSE	RMSE[a]	MAE[a]	FS	r
GRNN13[b]	2005	0.99	100%	−0.057	0.007	3.96	−2.60	−0.041	0.94
	2005–06	0.96	89%	−0.049	0.023	7.00	−1.07	0.041	0.91
GRNN10[c]	2005	0.99	100%	−0.054	0.007	3.99	−2.44	−0.053	0.94
	2005–06	0.98	91%	−0.056	0.015	5.71	−1.25	−0.025	0.94
GRNN8[d]	2005	0.99	100%	−0.059	0.008	4.32	−2.68	−0.059	0.94
	2005–06	0.97	89%	−0.062	0.016	5.98	−1.39	−0.034	0.94
PCR10[e]	2005	0.69	39%	−0.114	0.119	16.24	−5.37	0.496	0.55
	2005–06	0.73	39%	−0.118	0.113	15.57	−2.64	0.372	0.60

[a] RMSE and MAE values are presented in kg pc.
[b] Non-optimized GRNN model.
[c] GRNN model with inputs selected by ISF elimination procedure.
[d] GRNN model with inputs selected by ISF add-on procedure.
[e] Conventional model with the best results.

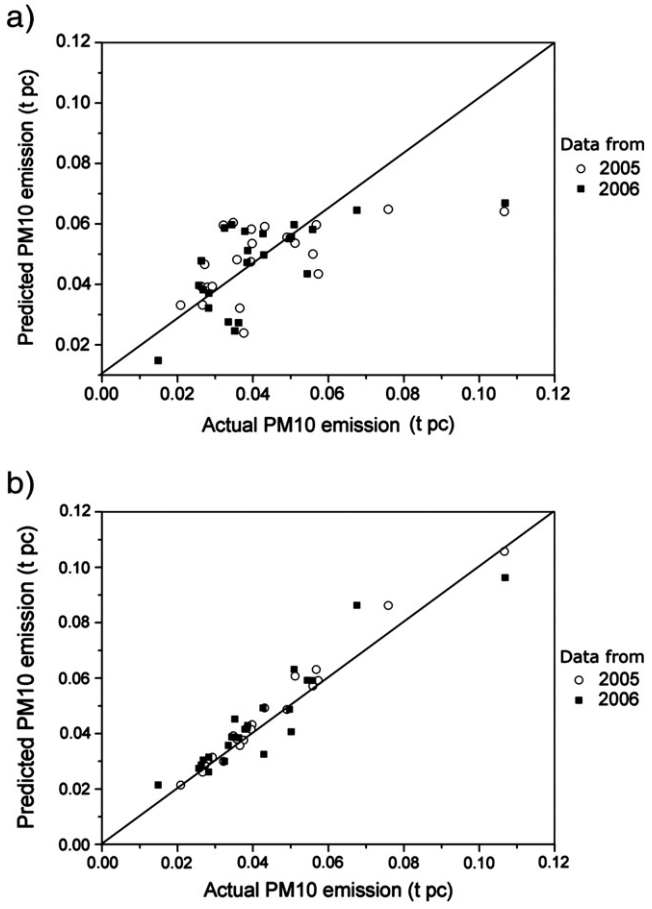

Fig. 4. Comparison of the "actual" data and model results for PM_{10} emission of European countries a) PCR10 b) GRNN10.

5.3. Comparison with conventional models

In order to investigate if input variable optimization, has any effect on the conventional models, two MLR and two PCR models were created:

- MLR13 and PCR13 — models created with all available inputs (Table 1);
- MLR10 and PCR10 — models created with inputs as the GRNN10 model (Table 3).

The MLR and PCR models were created with data from the training data set. For the PCR model creation, 24 randomly chosen patterns were used as validation data (the same numbers of patterns were used as the validation set for GRNN model training).

When the MLR and PCR models were applied to the training data set, MLR13 and PCR13 had better performance than the models with optimized inputs (MLR10 and PCR10). But when the real forecast capabilities of the created MLR and PCR models were tested, with data from test dataset, results were significantly different. The conventional models with optimized inputs had up to 50% better *IA* than the non-optimized models. The PCR10 model has slightly better results than the MLR10 model.

Nevertheless, comparison of the GRNN10 model performance with the performance of PCR10 model (Table 4) clearly shows that the GRNN model has much better forecast capabilities. Fig. 4a and b shows the performance of the GRNN10 and PCR10 models on the test dataset by comparing the results given by the models and the actual PM_{10} emission, respectively.

For data from year 2005, prediction results of the GRNN10 model were within the *FA1.25*. For data from year 2006, the GRNN10 model had an overestimated PM_{10} emission beyond *FA1.25* in cases of Luxembourg, Slovenia, Estonia and was underestimated in the case of Romania.

In some cases, the deviation that the model predicted from the official PM_{10} emission values can be related to the quality of input and output data. For example, the biggest relative error of 43% GRNN model was made in the case of Luxembourg. The Eurostat (2011) data shows that a decrease in the PM_{10} emissions of Luxembourg in year 2006, compared with the previous year, was 28% (unusually high), whilst the decrease in PM_{10} emissions in 2005 was only 2.6%. For 2005, in the case of Luxembourg the GRNN model produced a relative error of only 2.6%, while for the year 2006 the relative error was fourteen times bigger. It should be noted that Luxembourg is a very specific case, with size and economy structure significantly different from most of the other EU countries. Therefore Luxembourg was represented in the training data set with only 4% per country of the total

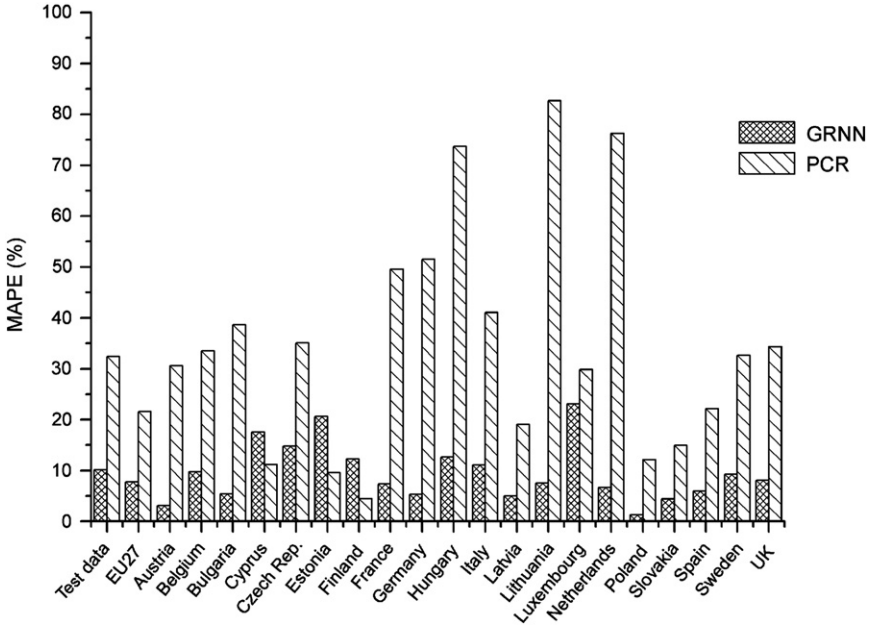

Fig. 5. Values of MAPE for test dataset, EU27 and EU countries.

learning patterns. This fact significantly affects the ability of the GRNN model to give accurate emission predictions.

In the case of Slovenia and Estonia the Eurostat (2011) data shows that the decrease in PM_{10} emissions in year 2006 compared with 2004 was 22.1% and 21.6%, respectively (Table 2). These are unusually high PM_{10} emission decreases, compared with most other countries (Table 2).

The values of mean absolute percentage error (*MAPE*) for predictions made by the GRNN10 and PCR10 models on the test dataset for EU countries are presented in Fig. 5. Only the countries for which there was data from all two test years are presented in Fig. 5.

The GRNN model has given better forecast results for the overall test dataset, EU27 and in case of most of the EU countries (Fig. 5). Only in the case of Cyprus, Estonia and Finland, the PCR10 model has a prediction with smaller *MAPE* values than the GRNN10 model.

6. Conclusions

This paper presented the development of an artificial neural network (ANN) model based on sustainability, economical and industrial parameters for the prediction of annual PM_{10} emission at the national level, and it's benchmarking versus conventional models.

The ANN model performance, after the optimization inputs and training parameters, was improved significantly in comparison with non-optimized ANN model, which included all available input variables. It can be concluded that ANN input selection, based on the individual smoothing factor values, can be successfully applied, resulting in more accurate models with less inputs needed.

A comparison of emission inventory based approaches and the ANN approach to PM_{10} emission modeling showed that the ANN model uses only a small fraction of the performance indicators needed by emission inventory based models. The fact that emission factors have not been used for the creation of ANN model significantly reduces the number of input parameters needed, the time needed for data preparation and the uncertainties when the model is used for future emission predictions with forecasted input variables. Also, a comparison of the optimized ANN model with conventional multi-linear regression (MLR) and principal component regression (PCR) shows that the ANN model has much better forecast performance than the MLR and PCR models when all models were trained and tested using the same datasets and input variables.

The presented ANN model can be used not only for PM_{10} emission forecasting, but also for simulating various scenarios of PM_{10} emission by changing the values of the input variables, e.g. to simulate possible consequences of regulatory actions on PM_{10} emissions. The results obtained by simulating various PM_{10} emission scenarios can be used by regulatory bodies and governments to support policy creation and implementation process, and for the development of strategies for improving air quality at regional and national levels.

Further researches are planned in expanding the model to include other environmental quality indicators such as emission of ozone precursors and acid oxides, using the same global group of sustainability, economical and industrial parameters, and in applying new techniques for input optimization, such as principal component analysis and correlation analysis.

Acknowledgments

The authors are grateful to the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 172007 for financial support.

References

Abdul-Wahab SA, Bakheit CS, Al-Alawi SM. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ Modell Softw 2005;20:1263–71.

Al-Alawi SM, Abdul-Wahab SA, Bakheit CS. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environ Modell Softw 2008;23:396–403.

Antanasijević D, Pocajt V, Popović I, Redžić N, Ristić M. The Forecasting of Municipal Waste Generation Using Artificial Neural Networks and Sustainability Indicators. Sustain Sci, in press. DOI: http://dx.doi.org/10.1007/s11625-012-0161-9.

Böhringer C, Jochem PEP. Measuring the immeasurable — a survey of sustainable indices. Ecol Econ 2007;63:1–8.

Ciancarella L, Briganti G, Calori G, Cappelletti A, Cionni I, Costa M, Cremona G, D'Elia I, D'Isidoro M, Finardi S, Mauri L, Mircea M, Pace G, Piersanti A, Racalbuto S, Radice P, Righini G, Vialetto G, Vitali L, Zanini G. National Italian integrated atmospheric model on air pollution: sensitivity to emission inventory. 14th conference on harmonisation within atmospheric dispersion modelling for regulatory purposes — 2–6 October 2011, Kos, Greece; 2011.

Dennis R, Fox T, Fuentes M, Gilliland A, Hanna S, Hogrefe C, et al. A framework for evaluating regional-scale numerical photochemical modeling systems. Environ Fluid Mech 2010;10:471–89.

European Commission. Council Directive 96/61/EC of 24 September 1996 concerning integrated pollution prevention and control. OJEC 1996;257:26–40.

European Commission. Directive 98/69/EC of the European Parliament and of the Council of 13 October 1998 relating to measures to be taken against air pollution by emissions from motor vehicles and amending Council Directive 70/220/EEC. Official Journal of the European Communities 1998;350:1-65.

European Commission. Council Directive 1999/30/EC of 22 April 1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air. Official Journal of the European Communities 1999;163: 41–60.

European Commission. Directive 2001/81/EC of the European Parliament and of the Council of 23 October 2001 on national emission ceilings for certain atmospheric pollutants. Official Journal of the European Communities 2001;309:22–30.

European Environment Agency (EEA). EMEP/EEA emission inventory guidebook — 2009. http://www.eea.europa.eu/publications/emep-eea-emission-inventory-guidebook-2009, 2009.

European Environment Agency (EEA). http://www.eea.europa.eu/data-and-maps/indicators/emissions-of-primary-particulates-outlook#data_specifications 2012.

Eurostat. http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database 2011.

Eurostat. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Air_pollution_statistics 2012.

GAINS EUROPE. RAINS review 2004. http://gains.iiasa.ac.at/index.php/documentation-of-model-methodology/model-reviews/rains-review-2004 2012.

GAINS EUROPE. http://gains.iiasa.ac.at/index.php/documentation-of-model-methodology/model-reviews/gains-review-2009 2012.

Grivas G, Chaloulakou A. Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens Greece. Atmos Environ 2006;40: 1216–29.

Hanna SR, Lu Z, Frey HC, Wheeler N, Vukovich J, Arunachalam S, et al. Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. Atmos Environ 2001;35:891–903.

International Fertilizer Industry Association (IFA). http://www.fertilizer.org/ifa/ifadata/search 2011.

International Iron and Steel Institute (IISI). Steel Statistical Yearbook 1999–2006. http://www.worldsteel.org/statistics/statistics-archive.html 2011.

Jakeman AJ, Letcher RA, Norton JP. Ten iterative steps in development and evaluation of environmental models. Environ Modell Softw 2006;21:602–14.

Kalogirou AS. Artificial intelligence for the modeling and control of combustion processes: a review. Prog Energ Combust 2003;29:515–66.

Karacan CÖ. Modeling and prediction of ventilation methane emissions of U.S. longwall mines using supervised artificial neural networks. Int J Coal Geol 2008;73:371–87.

Kassomenos P, Karakitsios S, Papaloukas C. Estimation of daily traffic emissions in a South-European urban agglomeration during a workday. Evaluation of several "what if" scenarios. Sci Total Environ 2006;370:480–90.

Kim S, Kim HS. Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. J Hydrol 2008;351:299–317.

Koo Y-S, Kim S-T, Cho J-S, Jang Y-K. Performance evaluation of the updated air quality forecasting system for Seoul predicting PM_{10}. Atmos Environ 2012;58:56–69.

Lim Y, Moon YS, Kim TW. Artificial neural network approach for prediction of ammonia emission from field-applied manure and relative significance assessment of ammonia emission factors. Eur J Agron 2007;26:425–34.

Najafi G, Ghobadian B, Tavakoli T, Buttsworth DR, Yusaf TF, Faizollahnejad M. Performance and exhaust emissions of a gasoline engine with ethanol blended gasoline fuels using artificial neural network. Appl Energy 2009;86:630–9.

Palani S, Liong SY, Tkalich P. An ANN aplication for water quality forecasting. Mar Pollut Bull 2008;56:1586–97.

Patel VC, Kumar A. Evaluation of three air dispersion models: ISCST2, ISCLT2, and SCREEN2 for mercury emissions in an urban area. Environ Monit Assess 1998;53:259–77.

Pay MT, Jiménez-Guerrero P, Baldasano JM. Assessing sensitivity regimes of secondary inorganic aerosol formation in Europe with the CALIOPE-EU modeling system. Atmos Environ 2012;51:146–64.

Radojević DM, Pocajt VV, Popović IG, Perić-Grujić AA, Ristić MDJ. Forecasting of greenhouse gas emissions in Serbia using Artificial neural networks. Energ Source Part A, in press. DOI http://dx.doi.org/10.1080/15567036.2010.514597.

Russell A, Dennis R. NARSTO critical review of photochemical models and modeling. Atmos Environ 2000;34:2283–324.

Scapellato ML, Canova C, Simone A, Carrieri M, Maestrelli P, Simonato L, et al. Personal PM_{10} exposure in asthmatic adults in Padova, Italy: seasonal variability and factors

affecting individual concentrations of particulate matter. Int J Hyg Environ Health 2009;212:626–36.
Schöpp W, Amann M, Cofala J, Heyes C, Klimont Z. Integrated assessment of European air pollution emission control strategies. Environ Modell Softw 1999;14:1–9.
Schöpp W, Klimont Z, Suutari R, Cofala J. Uncertainty analysis of emission estimates in the RAINS integrated assessment model. Environ Sci Policy 2005;8:601–13.
Sfetsos A, Vlachogiannis D. A new methodology development for the regulatory forecasting of PM_{10}. Application in the Greater Athens Area, Greece. Atmos Environ 2010;44: 3159–72.
Sözen A, Gülseven Z, Arcaklioğlu E. Estimation of GHG emissions in Turkey using energy and economic indicators. Energy Source Part A 2009;31:1141–59.
Specht DF. A general regression neural network. IEEE Trans Neural Netw 1991;2: 568–76.
U.S. Environmental Protection Agency (US EPA). AP 42. Compilation of air pollutant emission factors, I. Stationary point and area sources, Fifth edition; 1995. http://www.epa.gov/ttnchie1/ap42/.
U.S. Geological Survey (USGS). U.S. Geological Survey Minerals Yearbook—1999–2006, III; 2011. http://minerals.usgs.gov/minerals/pubs/country/europe.html,.
Voukantsis D, Karatzas K, Kukkonen J, Räsänen T, Karppinen A, Kolehmainen M. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 2011;409:1266–76.