International Journal of Greenhouse Gas Control 20 (2014) 244–253

Contents lists available at ScienceDirect

International Journal of Greenhouse Gas Control

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

Forecasting GHG emissions using an optimized artificial neural network model based on correlation and principal component analysis

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

ARTICLE INFO

Article history:
Received 13 May 2013
Received in revised form
27 September 2013
Accepted 15 November 2013
Available online 5 December 2013

Keywords:
General Regression Neural Network
GHG emission forecasting
Principal component analysis
Principal component regression

ABSTRACT

The prediction of GHG emissions is very important due to their negative impacts on climate and global warming. The aim of this study was to develop a model for GHG forecasting emissions at the national level using a new approach based on artificial neural networks (ANN) and broadly available sustainability, economical and industrial indicators acting as inputs. The ANN model architecture and training parameters were optimized, with inputs being selected using correlation analysis and principal component analysis. The developed ANN models were compared with the corresponding multiple linear regression (MLR) model, while an ANN model created using transformed inputs (principal components) was compared with a principal component regression (PCR) model. Since the best results were obtained with the ANN model based on correlation analysis, that particular model was selected for the actual 2011 GHG emissions forecasting. The relative errors of the 2010 GHG emissions predictions were used to adjust the ANN model predictions for 2011, which subsequently resulted in the adjusted 2011 predictions having a *MAPE* value of only 3.60%. Sensitivity analysis showed that gross inland energy consumption had the highest sensitivity to GHG emissions.

1. Introduction

Greenhouse gas (GHG) emissions and their impact on global warming have become a major concern, since global warming, and climate change in general are regarded as the most challenging problems facing the world today (Desjardins et al., 2007). In order to understand the human impact on the change in global climate, it is necessary to obtain reliable information on man-induced fluxes of greenhouse gases into (and from) the atmosphere (Winiwarter and Rypdal, 2001). The United Nation Framework Convention on Climate Change (UNFCCC) (UN, 1992) and the Kyoto Protocol are the first global efforts to mitigate GHG emissions. The Kyoto Protocol came into effect in 2005, and one of its primary objectives is the reduction of GHG emissions by 5.2% compared to the level in 1990, and with an aim to reach this level between the years 2008 and 2012. Large GHG emission reductions are also required for sustainable development and it will be difficult to achieve effective reductions that involve all countries without suitable methods and their associated models which are specifically developed to allow the simulation of a range of GHG emissions scenarios.

The main sources of GHG emissions data are emission inventories, which are a compilation of a large number of input parameters. The way these parameters have been processed to yield the final result, i.e. the total emission, depends on the emission model used. In general, most of emission sectors are estimated by multiplying the emission factor (*EF*) with the activity rate (*A*), a statistical parameter for the respective source. In practice, none of the input parameters (*EF* or *A*) is exactly known. In an emission inventory, the values of the parameters are determined as best "estimates" (Winiwarter and Rypdal, 2001) and further details on GHG emission inventories and related uncertainties can generally be found elsewhere (Rypdal and Winiwarter, 2001; Winiwarter and Rypdal, 2001; Monni et al., 2004; Wang and Chen, 2012).

GHG emission estimations for different emission sectors were also the subject of many studies in which predictions were obtained using different modeling approaches (Hediger, 2006; Dornburg et al., 2007; Chicco and Mancarella, 2008; Mancarella and Chicco, 2008; Syri et al., 2008; Matsumoto, 2008; Akimoto et al., 2010; Villalba and Gemechu, 2011; Couth et al., 2011; Rentziou et al., 2012). One of the most significant predictive models is the Greenhouse Gas and Air Pollution Interactions and Synergies (GAINS)

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

1750-5836/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijggc.2013.11.011

model (GAINS EUROPE, 2013) which estimates current and future emissions based on activity data, uncontrolled emission factors, the removal efficiency of emission control measures and the extent to which such measures are applied (Amann et al., 2011):

$$E_{i,p} = \sum_{k}\sum_{m} A_{i,k} ef_{i,k,m,p} x_{i,k,m,p} \quad (1)$$

where: i, k, m, p-Country, activity type, abatement measure, pollutant, respectively, $E_{i,p}$, emissions of pollutant p in country i, $A_{i,k}$, activity level of type k in country i, $ef_{i,k,m,p}$, emission factor of the pollutant p for the activity k in country i after the application of control measure m, $x_{i,k,m,p}$, share of total activity of type k in country i to which a control measure m for pollutant p is applied. Therefore GAINS can be regarded as an "upgraded" inventory approach for pollutant emission estimations.

An artificial neural network (ANN) is a non-linear computing system consisting of a large number of interconnected processing units (neurons), which simulates human brain learning (Balas et al., 2010). Neurons are linked together by synapses and organized in layers. The neural network architecture defines its structure including the number of layers, the number of neurons per layer, the learning algorithm, etc. Haykin (1994) describes a neural network as a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. Zhang et al. (1998) stated distinguishing features of ANNs, which make them valuable for forecasting:

- ANNs are data-driven self-adaptive methods – they learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe, as opposed to the traditional model-based methods, and can be regarded as the multivariate nonlinear nonparametric statistical method,
- after the learning phase ANNs can generalize – forecasting is performed via the prediction of future behavior (the unseen part) from examples of past behavior,
- ANNs are universal functional approximators – it has been shown that an ANN can approximate any continuous function to any desired accuracy,
- ANNs are nonlinear – the traditional approaches to time series prediction, such as the Box–Jenkins (Box and Jenkins, 1976) or ARIMA (Zhang, 2003) method, assume that the time series under study are generated from linear processes and they are totally inappropriate if the underlying mechanism is nonlinear.

ANNs have been successively applied for predicting GHG emissions based on sectoral energy consumption in Turkey (Sözen et al., 2007), as well as for GHG emissions prediction for some other European countries (Sözen et al., 2009; Radojević et al., 2013).

The main difference between conventional inventory-based models (e.g. GAINS) and the ANN approach is that the ANN model is less complex, requires a smaller number of input parameters (usually up to 10), and, most importantly, inputs are not predetermined. Therefore, the ANN approach can be implemented on a case-by-case basis, with GHG emissions being predicted using the inputs available for each country. Such variants of the ANN model can be applied to GHG emission estimations whenever countries adequately predict input parameters needed for the models based on activity levels and emission factors.

The aim of this paper is to describe an optimized ANN model for predicting GHG emissions at the national level, created using broadly available sustainability, economical and industrial indicators. The selection process of input variables based on correlation and principal component analysis is outlined in detail, together with the results and the sensitivity of the results to the input data. After this introduction, the paper provides a materials and methods chapter, with information focused on data collection and processing, along with the details of the ANN architecture used and its optimization. In the results and discussion section, the performance metrics are described first, followed by the results of the created ANN models with all available inputs, and the selection of inputs based on the correlation analysis (CA) and principal component analysis (PCA). The corresponding multi-linear regression model (MLR) and principal component regression (PCR) models created for comparison with the ANN models are briefly described, with an analysis of the model performance on GHG emissions prediction for the year 2011 with evaluation using EEA (2013) data is presented at the end of the results and discussion section. The sensitivities of the best ANN model inputs are assessed using individual smoothing factors (ISFs), which were determined during the ANN model training by a genetic algorithm (GA) for each of the inputs. The conclusions of this paper summarize the ANN modeling results and approach as an alternative method for the prediction of GHG emissions for both developing and developed countries.

2. Materials and methods

2.1. Data collection and processing

European GHG emissions can be broken down by the economic activities that lead to their production: energy supply and use, transportation, agriculture, industrial processes and waste. In order to create a suitable ANN prediction model, selected input indicators need to cover all GHG emission sectors. Fig. 1 presents the share contribution of each sector in Europe's GHG emissions (EEA, 2013) as well as the available sustainability, economical and industrial indicators which have been selected as inputs.

The inputs and GHG emissions data were obtained from Eurostat (2013) and the United Nations Economic Commission for Europe (UNECE) (2013) databases. Eurostat reports the number of passenger cars by age in the following intervals: the number of cars younger than 2 years (N_{0-2}), the number of cars 2–5 years old (N_{2-5}), the number of cars 5 to 10 years old (N_{5-10}) and the number of cars older than 10 years old (N_{10+}). In order to scale these four reports into one indicator (age of the passengers cars – APC) we made an assumption that the average car age in the first three groups is the mean of the overall age interval of each group. Regarding the group of cars older than 10 years, we applied the general assumption that most of the cars are between 10 and 15 years old, however since it is known that a number of the cars are even older than 15 years, we adjusted the estimate and set the average car age at 13 years old. Therefore APC was calculated according to Eq. (2).

$$APC = \frac{N_{0-2} + 3.5 \cdot N_{2-5} + 7.5 \cdot N_{5-10} + 13 \cdot N_{10+}}{N_{0-2} + N_{2-5} + N_{5-10} + N_{10+}} \quad (2)$$

The model was trained, validated and tested with the data for 28 European countries for the period 2004–2010. The selected input variables were normalized per capita and/or per GDP value of EU27 in order to allow comparison of countries of different sizes. Then, this dataset was divided into two subsets: training set (data from 2004–2009) and test set (data from 2010). Descriptive statistics of the training data set are presented in Table 1.

2.2. Development of ANN models

The ANN architecture used in this study is the General Regression Neural Network (GRNN) (Specht, 1991), which has already demonstrated good results in environmental modeling (Palani et al., 2008; Antanasijević et al., 2013a,b,c). GRNN models were created using Neuroshell 2 software (Ward systems group, 1993).

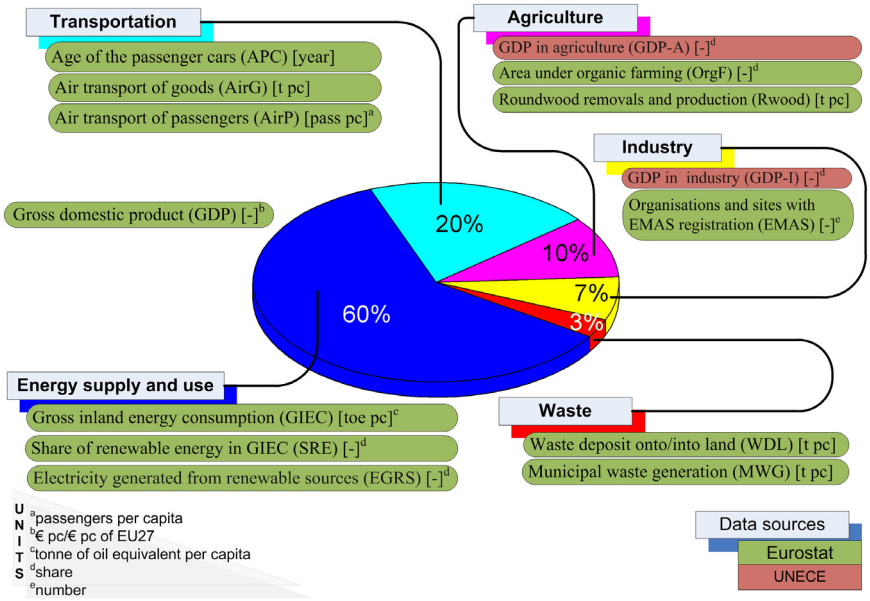

Fig. 1. GHG emission sources and selected model input indicators.

Table 1
Descriptive statistics of the training data set.

Indicator	Mean	St. Dev.	Min	Max
WDL [t pc]	0.25	0.16	0.002	0.68
SRE	0.14	0.14	0.001	0.65
Rwood [t pc]	1.64	2.36	0.000	10.88
EMAS	133.4	328.2	0.000	1641
MWG [t pc]	0.51	0.12	0.256	0.79
GIEC [toe pc]	3.91	1.68	1.653	10.34
GDP	1.01	0.45	0.341	2.74
EGRS	0.19	0.22	0.000	1.09
OrgF	0.05	0.04	0.000	0.19
AirP [pass pc]	2.55	2.10	0.147	9.10
AirG [t pc]	0.07	0.26	0.000	1.61
GDP-A	0.03	0.02	0.003	0.14
GDP-I	0.21	0.06	0.069	0.40
APC [year]	8.37	1.62	5.290	12.39
GHG [t pc]	10.93	4.03	4.811	27.84

GRNN is a one-pass supervised learning network consisting of four layers, and it uses a probability density function for the estimation of continuous variables. The number of neurons in the GRNN layers is defined by the number of input/output variables and the number of training cases used for model training:

- input layer has one neuron for each input variable,
- hidden layer has one neuron for each case in the training data set,
- summation layer has one neuron for each output variable, plus one,
- output layer has one neuron for each output variable.

The GRNN training process can be divided into two stages: first, the scaled values of input and output variables are presented to the hidden neurons through the input neurons and after that, the smoothing factor (SF) is determined. The GRNN architecture and training parameters along with the data flow are presented in Fig. 2.

The presented GRNN architectural and training parameters (Fig. 2) were varied during the GRNN optimization, in order to investigate the particular influence of each of these parameters on the overall performance of the GRNN for GHG emissions forecasting. As seen in Fig. 2, two different methods for validation data selection were tested: random method (12% of training data) and, since this is time-series data, data from the last training year (2009). Three different scale functions, which represent links between the input and the hidden neurons, were also tested (Eqs. (3)–(5)), along with two different distance metrics (Eqs. (6)–(7)):

$$\text{The linear scale function}: \quad f(x) = x \tag{3}$$

$$\text{The sigmoid (logistic) scale function}: \quad f(x) = \frac{1}{1+\exp(-x)} \tag{4}$$

$$\text{The hyperbolic-tangent (tanh) scale function}: \quad f(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} \tag{5}$$

$$\text{The Euclidean distance between two points, } a \text{ and } b, \text{ with } j \text{ dimensions}: \quad \sqrt{\sum_{i=1}^{j}(a_j - b_j)^2} \tag{6}$$

$$\text{The City block } (L_1;\ \text{Manhattan}) \text{ distance between two points, } a \text{ and } b, \text{ with } j \text{ dimensions}: \quad \sum_{i=1}^{j}\left|a_j - b_j\right| \tag{7}$$

Finally, the smoothing factor was determined using an iterative and genetic algorithm. The smoothing factor represents the width of the calculated Gaussian curve for each probability density function and it is a crucial parameter since it determines the accuracy of the GRNN. The smoothing factor must be greater than 0 and can usually range from 0.01 to 1 with reliable results (Thwin and Quah, 2005). Since the GRNN is a one-pass learning network, there is no problem with network overtraining, which can be characteristic of some other ANN architectures, but depending on the determined SF value, a model can over-fit or under-fit data (Fig. 3). This over-fitting can be interpreted as overtraining, and if a GRNN model over-fits the training data it will have a reduced generalization performance.

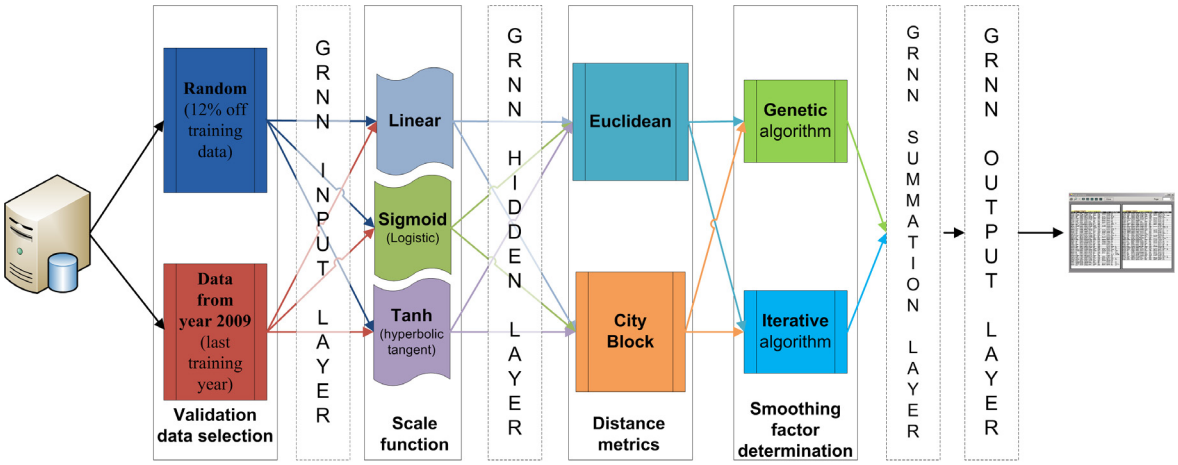

Fig. 2. GRNN architecture and training parameters with data flow.

3. Results and discussion

3.1. Performance metrics

All created GRNN models were analyzed using multiple performance indicators:

$$\text{The root mean squared error } (RMSE): \quad RMSE = \left[\overline{(C_p - C_o)^2}\right]^{1/2} \tag{8}$$

$$\text{The mean absolute error } (MAE): \quad MAE = \frac{1}{n}\sum\left|C_p - C_o\right| \tag{9}$$

$$\text{The index of agreement } (IA): \quad IA = 1\frac{\overline{(C_p - C_o)^2}}{\left[\left|C_p - \overline{C_o}\right| + \left|C_o - \overline{C_o}\right|\right]^2} \tag{10}$$

$$\text{The percent of predictions within a factor of 1.1 of the observed values } (FA1.1): \quad 0.9 < \frac{C_p}{C_o} < 1.1 \tag{11}$$

MAE and *RMSE* measure residual errors, which give a global understanding of the difference between the observed and modeled values. *IA* is a relative and bounded measure that allows for cross comparisons between models, and is limited to the range of 0–1 (Palani et al., 2008). *FA1.1* gives the percentage of cases in which the values of the ratio between observed and predicted concentration are in the range of 0.9–1.1. *FA1.1* is a very important performance indicator since it shows if the model has the same accuracy for every country for which predictions are made. Because of its significance we decided to select more rigid FA boundaries, in comparison with our previous studies (Antanasijević et al., 2013b,c) in which we used *FA1.25* (the percent of predictions within the factor of 1.25 of the observed values).

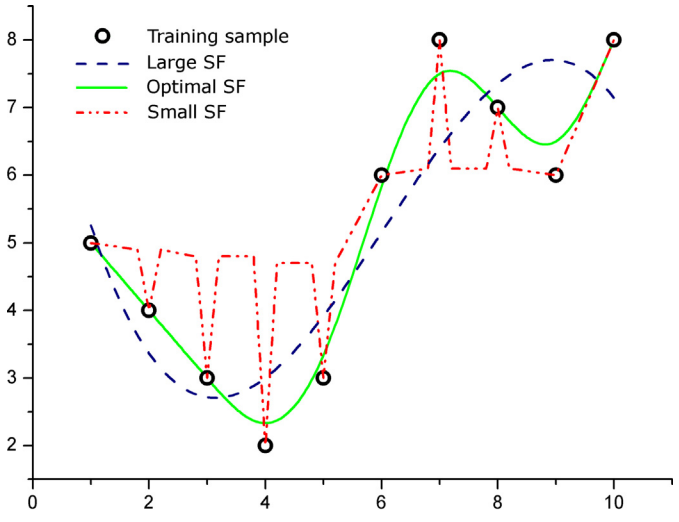

Fig. 3. Theoretical shape of GRNN fit of training data depending on different value of smoothing factor.

3.2. GRNN architecture and training parameters optimization

During the optimization of the architecture and training parameters, 24 GRNN models with different architecture and training parameter sets were created using Neuroshell 2 software. All tested GRNN models had 14 input neurons, 168 hidden, 2 summation and 1 output neuron (details on neuron numbers per layer are summarized in Section 2.2). The performance metric values for GRNN models with different parameters are presented in Table 2. GRNN models that had random validation data selection and SF determination by genetic algorithm performed significantly better, while models with different scale functions and distance metrics displayed a very similar performance.

The performance metric values for all created GRNN models during architecture and training parameters optimization are presented in Table 3. It can be observed that the GRNN model with the best performance is the GRNN3 model, which has random validation data selection, linear scale function, Euclidian distance metrics and a genetic algorithm for the determination of the smoothing factor.

Only the GRNN5 model has better values for one of the performance metrics (*FA1.1*) in comparison with model GRNN3.This means that the GRNN3 model made one more prediction with a relative error bigger than 10% in comparison with the GRNN5 model. The GRNN3 model can nevertheless be regarded as the best optimized model, with an opportunity to further improve performance through input variable selection (Section 3.3). Regarding the results in Table 3, it can be also concluded that tanh function performs better with an iterative algorithm, while sigmoid functions perform more effectively with a genetic algorithm.

Table 2
Performance indicators values for GRNN models with different parameters.

GRNN parameter	Performance indicators for test set			
	IA	*FA1,1* [%]	*MAE* [t pc]	*RMSE* [t pc]
Validation data selection	Average for 12 models			
Random selection	**0.991**	**88.1**	**0.465**	**0.701**
Data from 2009	0.985	68.9	0.753	0.930
Scale function	Average for 8 models			
Linear	0.987	78.5	0.611	0.812
Sigmoid	0.989	78.6	0.615	0.805
Tanh	0.988	78.4	0.602	0.831
Distance metrics	Average for 12 models			
Euclidean	0.988	79.4	0.607	0.813
City Block	0.988	77.6	0.611	0.818
SF determination algorithm	Average for 12 models			
Genetic	**0.989**	**80.5**	**0.595**	**0.787**
Iterative	0.987	76.5	0.624	0.845

Bold marked values emphasize the models with the best results.

A comparison of the GRNN3 results with a multi-linear regression model (MLR) created and tested using the same dataset showed that the GRNN3 model has demonstrated considerably better forecast performance (Table 2). The MLR model for GHG emissions prediction was formed using SPSS 19 software (IBM, 2010). The comparison of the GRNN3 and MLR model results with actual GHG emissions data for the test dataset are shown in Fig. 4.

3.3. Input variables selection

3.3.1. Correlation analysis

The performance of a GRNN model, as well as any general ANN model, depends on data representation (Cherkassky and Lari-Najafi, 1992) and the number of input variables, with too many inputs causing poor generalization performance (Tripathy, 2010). An important characteristic of data representation is whether or not input variables are correlated, since correlated data can introduce confusion to the neural network during the learning process (Walczak and Cerpa, 1999). Correlation problems between input variables can be solved by using correlation analysis (Antanasijević et al., 2013c) or principal component analysis (Balas et al., 2010; He and Ma, 2010).

Based on the correlation analysis (Table 4) two different input datasets were defined: the "independent input variables" dataset that contains only variables with a mutual coefficient of correlation, less than 0.80 and the "input-output variables correlated" dataset that contains only variables correlated with outputs with a coefficient of correlation higher than 0.10.

In the present case, there are two pairs of correlated inputs (with a mutual coefficient of correlation bigger than 0.80):

- gross domestic product (GDP) and gross inland energy consumption (GIEC)

Table 3
List of GRNN model created during the architecture and training parameters optimization with performance indicators values. The performance indicators values for MLR model are presented for comparison.

Model[a]	Validation data selection	Scale function	Distance metrics	SF determination algorithm	Performance metrics values for test set			
					IA	*FA1,1* [%]	*MAE* [t pc]	*RMSE* [t pc]
GRNN1	**Random**[b] (12% off training data)	**Linear**	Euclidean	Genetic	**0.995**	89	0.422	0.541
GRNN2				Iterative	0.989	86	0.447	0.750
GRNN3			**City Block**	**Genetic**	**0.995**	89	**0.399**	**0.521**
GRNN4				Iterative	0.990	89	0.450	0.735
GRNN5		Logistic	Euclidean	Genetic	0.993	**93**	0.469	0.627
GRNN6				Iterative	0.988	86	0.512	0.780
GRNN7			City Block	Genetic	0.993	89	0.466	0.636
GRNN8				Iterative	0.989	86	0.489	0.770
GRNN9		Tanh	Euclidean	Genetic	0.988	86	0.501	0.793
GRNN10				Iterative	0.989	89	0.463	0.750
GRNN11			City Block	Genetic	0.989	86	0.489	0.761
GRNN12				Iterative	0.989	89	0.476	0.753
GRNN13	Data from year 2009	Linear	Euclidean	Genetic	0.983	75	0.778	0.999
GRNN14				Iterative	0.987	68	0.748	0.853
GRNN15			City Block	Genetic	0.975	71	0.862	1.183
GRNN16				Iterative	0.985	61	0.784	0.910
GRNN17		Logistic	Euclidean	Genetic	0.987	71	0.729	0.889
GRNN18				Iterative	0.985	68	0.775	0.959
GRNN19			City Block	Genetic	0.989	75	0.669	0.803
GRNN20				Iterative	0.984	61	0.807	0.978
GRNN21		Tanh	Euclidean	Genetic	0.987	71	0.692	0.877
GRNN22				Iterative	0.985	71	0.753	0.941
GRNN23			City Block	Genetic	0.989	71	0.659	0.810
GRNN24				Iterative	0.985	64	0.780	0.960
MLR	–	–	–	–	0.950	43	1.343	1.660

Bold marked values emphasize the models with the best results.
[a] All tested modes had 14 input, 168 hidden, 2 summation and 1 output neuron.
[b] Data from years 2004–2009.

Table 4
Correlation analysis results.

	WDL	SRE	RWood	EMAS	MWG	GIEC	GDP	EGRS	OrgF	AirP	AirG	GDPA	GDPI	APC
WDL[a]	1													
SRE	−0.35	1												
RWood	−0.21	0.62	1											
EMAS	−0.24	−0.09	−0.13	1										
MWG	0.08	−0.25	−0.26	0.20	1									
GIEC	−0.49	0.16	0.31	−0.03	0.30	1								
GDP	−0.41	0.10	−0.01	0.10	0.60	**0.84**	1							
EGRS	−0.40	**0.91**	0.38	0.01	−0.09	0.15	0.23	1						
OrgF	−0.37	0.43	0.41	0.18	−0.13	0.10	0.07	0.45	1					
AirP	0.34	−0.05	−0.16	−0.04	0.72	0.16	0.44	0.05	−0.18	1				
AirG	−0.16	−0.20	−0.10	−0.07	0.32	0.70	0.71	−0.16	−0.13	0.11	1			
GDPA	0.35	0.01	−0.02	−0.22	−0.46	−0.52	−0.67	−0.07	−0.25	−0.44	−0.27	1		
GDPI	−0.24	0.43	0.22	0.02	−0.45	−0.04	−0.19	0.42	0.11	−0.34	−0.39	0.20	1	
APC	0.31	0.19	0.21	−0.26	−0.64	−0.48	−0.69	−0.03	−0.06	−0.31	−0.37	0.54	0.25	1
GHG	**−0.15**	**−0.27**	−0.08	−0.04	**0.43**	**0.77**	**0.73**	**−0.23**	−0.06	**0.31**	**0.74**	**−0.41**	**−0.18**	**−0.47**

Bold marked values indicate the high correlated inputs and inputs significantly correlated with GHG emission.
[a] Input abbreviations are presented in Fig. 1.

- share of renewable energy in gross final energy consumption (SRE) and electricity generated from renewable sources (EGRS).

A correlation between these indicators was expected, since it is well known that energy consumption depends on the GDP value, and that the energy generated from renewable sources is primarily from electrical energy. Therefore, four GRNN models (marked as IV-GRNN3) with different combinations of independent input variables were created. Performance indicator values for the IV-GRNN3 models with corresponding combinations of independent input variables are presented in Table 5.

In addition, another GRNN model (CA-GRNN3) was created with a dataset that contains only variables correlated to GHG emissions with a coefficient of correlation higher than 0.10. As seen in Table 4, three initial inputs (area under organic farming (OrgF), roundwood removals and production (Rwood), plus organizations and sites with EMAS registration (EMAS)) had to be removed in order to create the "GHG correlated" input dataset. Performance metric values for CA-GRNN models are also presented in Table 5. All GRNN models created based on correlation analysis results used architecture and training parameter sets as with the GRNN3 model, only the number of input neurons had to be changed, since the number of input variables had been reduced.

The CA-GRNN3 model demonstrated the best values for all four performance metrics in comparison with the models with independent datasets (Table 5), and also much improved performance in comparison with the GRNN3 model created, using all available variables (Table 3). This proves that the reduction of input variables, particularly those which are not correlated with the output variables, significantly improves GRNN model performance. Performance metric values for the created CA-MLR model are also presented in Table 4, while the comparison of the CA-GRNN3 and CA-MLR model results with the actual GHG emissions data for the test dataset is presented in Fig. 5. It can be concluded that the CA-GRNN3 model has demonstrated much improved forecast performance in comparison with the corresponding CA-MLR model.

3.3.2. Principal component analysis

The PCA was performed using SPSS 19 software (IBM, 2010) on the reduced input data set (the same data set which was used for the CA-GRNN3 creation). First, it was necessary to verify that the PCA was applicable to this dataset by using the Kaiser–Meyer–Olkin (KMO) test of sampling adequacy and Bartlett's Sphericity Test. The Kaiser–Meyer–Olkin (KMO) measure is an index for comparing the magnitudes of observed correlation coefficients to the magnitudes of partial correlation coefficients. The Bartlett's Test of Sphericity examines the correlation matrix with a matrix of zero correlations

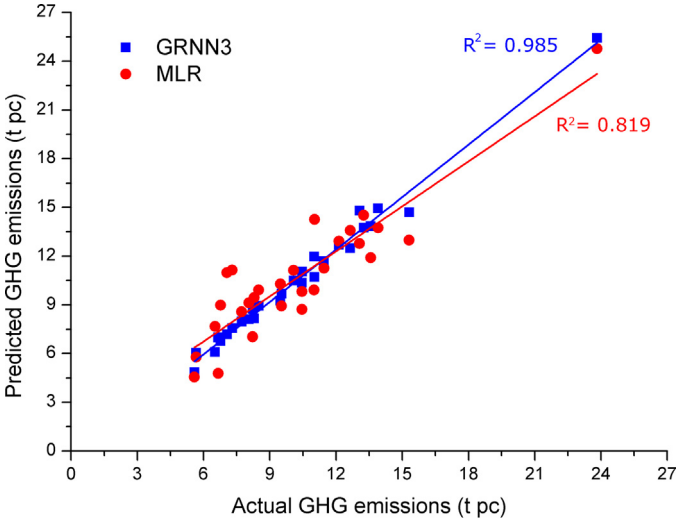

Fig. 4. Comparison of GRNN3 and MLR model results with actual GHG emissions data (test dataset).

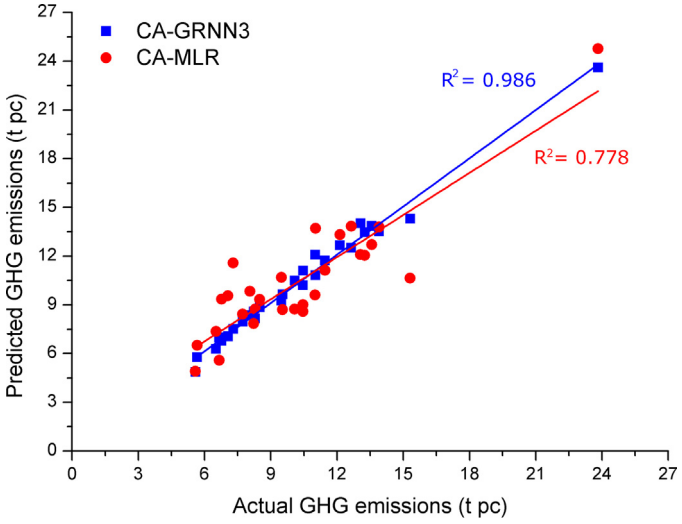

Fig. 5. Comparison of CA-GRNN3 and CA-MLR model results with actual GHG emissions data (test dataset).

Table 5
Performance indicators values for GRNN and MLR models created using variables selected based on correlation analysis results.

Model	Removed inputs	Performance indicators for test set			
		IA	*FA1,1* [%]	*MAE* [t pc]	*RMSE* [t pc]
IV1-GRNN3[a]	GDP and SRE	0.995	93	0.387	0.536
IV2-GRNN3	GDP and EGRS	0.994	**96**	0.433	0.557
IV3-GRNN3	GIEC and SRE	0.989	86	0.498	0.750
IV4-GRNN3	GiEC and EGRS	0.989	89	0.460	0.739
CA-GRNN3[b]	**Rwood, EMAS and OrgF**	**0.996**	**96**	**0.353**	**0.450**
CA-MLR	Rwood, EMAS and OrgF	0.940	46	1.401	1.759

Bold marked values emphasize the models with the best results.
[a] IV-GRNN3 models have 12 input neurons.
[b] CA-GRNN3 models has 11 input neurons.

(the identity matrix). A small p value indicates that the correlation matrix is not an identity matrix and therefore indicates that the PCA is appropriate. Kaiser (1974) recommends using KMO values greater than 0.5, which means that the KMO of 0.616 (Table 6) obtained in this study is acceptable. Barlett's test is highly significant ($p < 0.0001$) (Table 6) and therefore the PCA is appropriate for this data. Communality values, which range from 0 to 1 and indicate the amount of variability defined by PCs, are presented in Table 6. It can be seen that for almost all variables the extracted communality values were greater than 0.8.

Promax with Kaiser Normalization was used as a rotation method. The PC score coefficient matrix, eigenvalues and respective variances as well as total variance explained by extracted PC are shown in Table 7. The total variance explained by the five extracted PC's was almost 90%.

Table 6
List of inputs used in PCA. Communalities extracted, KMO and Bartlett's Test values.

Kaiser–Meyer–Olkin (KMO)		0.616
Bartlett's test	Approx. Chi-Square	2291.8
	df	55
	Sig. (p)	<0.0001

Input	Communalities extracted
SRE	0.965
MWG	0.842
GIEC	0.909
GDP	0.971
AirP	0.935
AirG	0.941
GDPA	0.700
APC	0.766
WDL	0.899
EGRS	0.947
GDPI	0.964

Table 7
The PC score coefficient matrix, eigenvalues and respective variances as well as total variance explained by extracted PC.

Input	Component				
	1	2	3	4	5
SRE	0.056	0.034	0.449	0.002	0.007
MWG	−0.185	−0.022	−0.027	0.293	−0.155
GIEC	−0.054	0.394	0.033	−0.012	0.202
GDP	−0.143	0.284	0.074	0.088	0.039
AirP	−0.011	0.006	0.041	0.547	0.038
AirG	0.110	0.461	−0.058	−0.024	−0.183
GDPA	0.321	0.012	0.001	−0.056	−0.090
APC	0.367	0.046	0.034	0.058	0.062
WDL	0.308	−0.005	−0.176	0.433	0.014
EGRS	−0.044	−0.033	0.452	0.012	−0.014
GDPI	−0.002	0.004	−0.002	0.005	0.876
Eigenvalues	4.258	2.714	1.469	0.843	0.553
Variance per comp. [%]	38.708	24.676	13.355	7.667	5.024
Cumulative variance [%]	38.708	63.384	76.739	84.406	89.43

Table 8
Performance indicators values for PCA-GRNN3 and PCR model.

Model	Performance indicators for test set			
	IA	*FA1,1* [%]	*MAE* [t pc]	*RMSE* [t pc]
PCA-GRNN3	0.986	89	0.637	0.942
PCR (PCA-MLR)	0.936	39	1.362	1.782

The fore mentioned five PCs were used as inputs for the PCA-GRNN3 model, which therefore had 5 input neurons, while other architectural and training parameters remained the same as those used in the GRNN3 model. For comparison, another MLR model based on extracted principal components was created, and since the combination of PCA with MLR is actually a principal component regression (PCR) model, the created PCA-MLR model will be referred to as the PCR model. Performance indicator values for PCA-GRNN3 and the PCR model is presented in Table 8, whilst the comparison of results of those models with the actual GHG emissions data for the test dataset are presented in Fig. 6.

The PCA-GRNN3 model demonstrated a slightly reduced generalization power in comparison with the CA-GRNN3 model (Table 5 and Fig. 5). Although some authors (Juhos et al., 2008) suggest that a reduction of dimensions can cause a loss of information during the process, which can subsequently result in a negative influence on the model performance (i.e. worse results than expected), however in the present case, based on the training result, it is more likely that the PCA-GRNN3 model was over trained (Section 2.2), and therefore over-fitted the data, which caused the apparent decrease in the PCA-GRNN3 model's generalization ability.

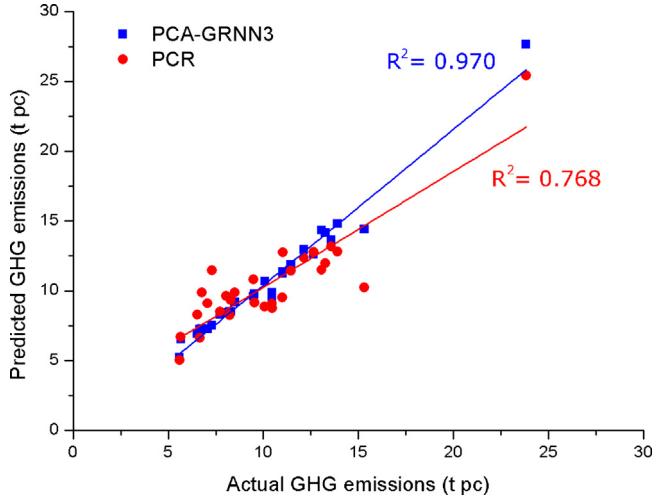

Fig. 6. Comparison of PCA-GRNN3 and PCR model results with actual GHG emissions data (test dataset).

Table 9
The CA-GRNN3 GHG emissions results for year 2010 (test data) with relative errors and predictions for year 2011.

Country	Year 2010			Year 2011		
	Actual values[a] [Mt]	Modeled values [Mt]	Rel. err. [%]	Actual values[c] [Mt]	Predictions [Mt]	Corrected[d] predictions [Mt]
Belgium	132.5	138.4	4.47	121.3	130.4	124.6
Bulgaria	61.4	64.1	4.37	67.9	65.8	62.9
Czech Rep.	139.2	141.3	1.53	141.1	136.2	134.1
Denmark	61.1	67.0	9.78	56.1	**64.5**	58.2
Germany	936.5	959.7	2.47	917	918.9	896.1
Estonia	20.5	19.2	−6.55	20.9	**18.7**	19.9
Ireland	61.3	62.6	2.13	57.3	**63.3**	62.0
Greece	118.3	125.6	6.15	118.5	125.8	118.0
Spain	355.9	367.6	3.28	356.1	369.3	357.2
France	522.4	541.7	3.69	497.5	533.5	513.8
Italy	501.3	494.4	−1.39	493.7	496.0	502.8
Cyprus	10.8	11.6	7.31	N/A	12.4	11.5
Latvia	12.1	**10.5**[b]	**−12.83**	12.1	**10.6**	12.0
Lithuania	20.8	20.1	−3.61	21.4	20.1	20.8
Luxembourg	12.1	12.0	−0.84	12.3	12.0	12.1
Hungary	67.7	67.9	0.31	65.6	67.7	67.5
Malta	3.0	3.1	3.02	N/A.	4.9	4.7
Netherlands	210.1	207.8	−1.08	195.8	207.6	209.8
Austria	84.6	88.0	4.02	81.9	88.3	84.8
Poland	400.9	391.8	−2.25	409.3	388.8	397.5
Portugal	70.6	74.2	5.13	70	74.0	70.2
Romania	121.4	123.9	2.10	123.7	125.5	122.9
Slovenia	19.5	19.8	1.31	19.5	20.8	20.5
Slovakia	46.0	47.8	3.99	45.9	45.7	43.9
Finland	74.6	72.6	−2.65	67.3	70.6	72.5
Sweden	66.2	66.2	−0.09	62.8	65.3	65.3
UK	590.2	576.9	−2.26	553.8	581.1	594.3
Norway	53.9	53.0	−1.74	52.7	53.2	54.1
Descriptive/ performance statistics[e]	MAPE[f] [%]	3.60			5.40	3.60
	Max rel. err. [%]	−12.83			15.01	8.20
	Min rel. err. [%]	−0.09			0.20	0.20
	FA1,1 [%]	96			85	100
	MAE [Mt]	4.7			6.8	6.0
	RMSE [Mt]	7.4			11.0	10.6

[a] Eurostat (2013).
[b] Bold marked values have rel. err. > 10%.
[c] EEA (2013).
[d] Correction is based on the relative errors of model predictions for year 2010.
[e] Cyprus and Malta were excluded from calculation for year 2011.
[f] The mean absolute percentage error $MAPE = 100\frac{1}{k}\sum\frac{|C_o - C_p|}{C_o}$.

3.4. GHG emissions predictions for the year 2011: error analysis and prediction

Since the CA-GRNN3 model has the best GHG emissions predictions for the test dataset, it was used for a GHG emissions estimation for 2011. The CA-GRNN3 model was able to provide a prediction for GHG emissions for 2011; since the GRNN model inputs values were available for the studied European countries (GHG emissions for 2011 were not available on Eurostat). Since the CA-GRNN3 model was tested using data from 2010, it could be assumed that the relative error for 2010 GHG emissions prediction per country will be similar to the relative error of 2011 GHG emissions predictions for the very same country. Due to this fact, the relative errors of the 2010 GHG emissions predictions were used to correct the CA-GRNN3 model predictions for 2011.

Actual and modeled GHG emissions for 2010 with the relative errors for each studied country along with descriptive and performance statistics are presented in Table 9. Only in the case of Latvia the CA-GRNN3 model made a relative error bigger than 10%. The mean absolute percentage error (MAPE) for all 2010 predictions was only 3.60% (Table 9).

The CA-GRNN3 prediction for 2011, with actual 2011 GHG emissions obtained from EEA (2013), for all countries accept Malta and Cyprus are also presented in Table 9. A comparison with actual GHG emissions for 2011 shows that the model made predictions with a maximum relative error of 15%, while in the case of four countries (Denmark, Estonia, Ireland and Latvia), the relative error was bigger than 10%. GHG emission predictions for 2011 with corrections based on the relative errors of CA-GRNN3 model predictions for the year 2010 are also presented in Table 9. All corrected predictions for 2011 have relative errors less than 10%, with a maximum relative error of 8%. Also, the MAPE value for all 2011 predictions was only 3.60%, the same as the data in 2010, which indicates that model error was well controlled.

A comparison of corrected GHG emissions predictions for 2011 obtained by the CA-GRNN3 model with the actual GHG emissions is presented in Fig. 7.

3.5. Sensitivity analysis

Sensitivity analysis provides an understanding of the significance of the model input variables. Sensitivity analysis in the case of the GRNN model which is trained using a genetic algorithm (GA) can be done by analyzing individual smoothing factors (ISFs) determined by the GA for each input. The ISF value ranges between 0 and 3: the larger the factor for a given input is, the more important the input is to the model (Antanasijević et al., 2013b). A sensitivity analysis (Table 10) was performed using ISFs determined during the best GRNN model (CA-GRNN3) training.

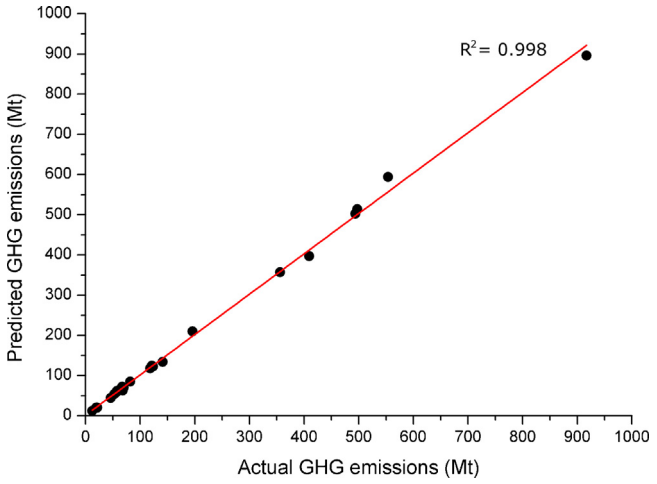

Fig. 7. Comparison of corrected GHG emissions predictions for 2011 obtained by CA-GRNN3 with actual GHG emissions.

Table 10
Individual smoothing factors determined during CA-GRNN3 model training.

Input	ISF
Gross inland energy consumption (GIEC)	3.00
Air transport of passengers (AirP)	2.07
Electricity generated from renewable sources (EGRS)	0.91
GDP in agriculture (GDP-A)	0.73
Share of renewable energy (SRE)	0.41
Municipal waste generation (MWG)	0.19
Air transport of goods (AirG)	0.05
Age of the passenger cars (APC)	0.05
Gross domestic product (GDP)	0.02
GDP in industry (GDP-I)	0.02
Waste deposit onto or into land (WDL)	0.02

It is evident that gross inland energy consumption (GIEC) had the highest impact on the output (ISF = 3), followed by air transport of passengers (AirP), electricity generated from renewable sources (EGRS), GDP in agriculture (GDP-A) and share of renewable energy (SRE). The output had relatively low sensitivity to municipal waste generation (MWG), air transport of goods (AirG), age of the passenger cars (APC), gross domestic product (GDP), GDP in industry (GDP-I) and waste deposit onto or into land (WDL). The fact that GDP has a surprisingly low significance on the model only confirms the correlation analysis results (Table 3), and means that the GRNN, among two high correlated inputs (GDP and GIEC), has "decided" to build a model based on GIEC which holds the highest correlation to GHG emissions.

4. Conclusions

The main goal of this study was to develop a model for GHG emissions forecasting for European countries using an Artificial Neural Network (ANN) approach with sustainability, economical and industrial indicators being used as inputs. The selected ANN architecture, General Regression Neural Network (GRNN), was optimized with regards to both its architecture and training parameters. The input selection was carried out using Correlation analysis and Principal Component Analysis (PCA). Available data for 28 European countries for the period 2004–2010 was used and the performance evaluation was done using multiple statistical performance indicators: *AI*, *FA1.1*, *RMSE*, *MAE* and *MAPE*.

The optimization of architecture and training parameters showed that the model called GRNN3 demonstrated the best results; this model had random validation data selection, linear scale function, Euclidian distance metrics and a genetic algorithm for determination of the smoothing factor. After the first selection was made, the results of different versions of the GRNN3 model were compared: a GRNN3 model with all available inputs, a model created based on correlation analysis (CA-GRNN3) and a model that used principal components as inputs (PCA-GRNN3). All three of these models were also compared with the corresponding multilinear regression model (MLR) and principal component regression (PCR) models.

Correlation analysis showed that for the CA-GRNN3 model, three of the fourteen inputs could be removed and so the PCA was subsequently applied to the revised and reduced input data set. Total variance explained by the five extracted PC's was almost 90%. Although the PCA-GRNN3 model has demonstrated better forecast performance in comparison with the corresponding PCR model, it has a slightly reduced generalization power in comparison with the CA-GRNN3 model. It can be concluded that the correlation analysis based selection of model inputs was the most successfully approach, resulting in more accurate models with less input parameters needed.

The CA-GRNN3 model was also applied for 2011 GHG emissions forecasting. In order to enhance forecast performance, the relative errors of the 2010 GHG emissions predictions were used to adjust the CA-GRNN3 model predictions for 2011. The adjusted GHG predictions for 2011 had a maximum relative error of 8%, with a *MAPE* of only 3.60%, the same as in 2010 (test data), which indicates that the model error was well controlled.

Finally, sensitivity analysis based on individual smoothing factors (ISFs) showed that gross inland energy consumption (GIEC) had the highest sensitivity on output and also holds the highest correlation to GHG emissions.

It can be concluded that the created ANN model can be successfully applied for GHG emissions forecasting at the national level, which is needed for sustainability reporting and for the analysis of fulfillment of the Kyoto Protocol objectives. The fact that inputs of an ANN GHG model are not pre-determined enables its application on a case-by-case basis, with GHG emissions being predicted using the available sustainability, economical and industrial indicators for each country. This is also a key advantage of the ANN approach in comparison with conventional emission inventory based models which commonly use strictly defined input parameters, the determination of which, require substantial and time consuming field studies. Furthermore, the ANN model allows simulation of different GHG emission scenarios by varying the values of input variables as a result of proposed GHG reduction strategies.

Further research is planned in expanding the model to include other environmental quality indicators such as emission of ozone precursors and acid oxides, as well as analyzing the contribution of various sectors of industry and transportation at a national and regional level.

Acknowledgements

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

Akimoto, K., Sano, F., Homma, T., Oda, J., Nagashima, M., Kii, M., 2010. Estimates of GHG emission reduction potential by country, sector, and cost. Energy Policy 38, 3384–3393.

Amann, M., Bertok, I., Borken-Kleefeld, J., Cofala, J., Heyes, C., Höglund-Isaksson, L., Klimont, Z., Nguyen, B., Posch, M., Rafaj, P., Sandler, R., Schöpp, W., Wagner, F., Winiwarter, W., 2011. Cost-effective control of air quality and greenhouse

gases in Europe: modeling and policy applications. Environ. Modell. Softw. 26, 1489–1501.
Antanasijević, D., Pocajt, V., Popović, I., Redžić, N., Ristić, M., 2013a. The forecasting of municipal waste generation using artificial neural networks and sustainability indicators. Sustain. Sci. 8, 37–46.
Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ., Perić-Grujić, A.A., 2013b. PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519.
Antanasijević, D., Ristić, M., Perić-Grujić, A., Pocajt, V., 2013c. Forecasting human exposure to PM10 at the national level using artificial neural network approach. J. Chemometr. 27, 170–177.
Balas, C.E., Koç, M.L., Tür, R., 2010. Artificial neural networks based on principal component analysis, fuzzy systems and fuzzy neural networks for preliminary design of rubble mound breakwaters. Appl. Ocean Res. 32, 425–433.
Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA.
Cherkassky, V., Lari-Najafi, H., 1992. Data representation for diagnostic neural networks. IEEE Intell. Syst. 7, 43–44, 47, 49, 51–53.
Chicco, G., Mancarella, P., 2008. Assessment of the greenhouse gas emissions from cogeneration and trigeneration systems. Part I: models and indicators. Energy 33, 410–417.
Couth, R., Trois, C., Vaughan-Jones, S., 2011. Modelling of greenhouse gas emissions from municipal solid waste disposal in Africa. Int. J. Greenh. Gas Control 5, 1443–1453.
Desjardins, R.L., Sivakumar, M.V.K., de Kimpe, C., 2007. The contribution of agriculture to the state of climate: workshop summary and recommendations. Agr. Forest Meteorol. 142, 314–324.
Dornburg, V., van Dam, J., Faaij, A., 2007. Estimating GHG emission mitigation supply curves of large-scale biomass use on a country level. Biomass Bioenergy 31, 46–65.
European Environment Agency (EEA), 2013. Greenhouse Gas Emission Trends and Projections in Europe 2012. http://www.eea.europa.eu/publications/ghg-trends-and-projections-2012
Eurostat, 2013. http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database
GAINS EUROPE, 2013. http://gains.iiasa.ac.at/index.php/documentation-of-model-methodology/model-reviews/gains-review-2009
Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Macmillan College Publishing, New York.
He, F., Ma, C., 2010. Modeling greenhouse air humidity by means of artificial neural network and principal component analysis. Comput. Electron. Agric. 71S, S19–S23.
Hediger, W., 2006. Modeling GHG emissions and carbon sequestration in Swiss agriculture: an integrated economic approach. Int. Congr. Ser. 1293, 86–95.
IBM Corp, 2010. IBM SPSS Statistics for Windows, Version 19. Armonk, NY, USA.
Juhos, I., Makra, L., Tóth, B., 2008. Forecasting of traffic origin NO and NO_2 concentrations by support vector machines and neural networks using principal component analysis. Simul. Modell. Pract. Theory 16, 1488–1502.
Kaiser, H.F., 1974. Index of factorial simplicity. Psychometrika 39, 31–36.
Mancarella, P., Chicco, G., 2008. Assessment of the greenhouse gas emissions from cogeneration and trigeneration systems. Part II: analysis techniques and application cases. Energy 33, 418–430.
Matsumoto, K., 2008. Evaluation of an artificial market approach for GHG emissions trading analysis. Simul. Model. Pract. Theory 16, 1312–1322.
Monni, S., Syri, S., Savolainen, I., 2004. Uncertainties in the Finnish greenhouse gas emission inventory. Environ. Sci. Policy 7, 87–98.
Palani, S., Liong, S.Y., Tkalich, P., 2008. An ANN aplication for water quality forcasting. Mar. Pollut. Bull. 56, 1586–1597.
Radojević, D.M., Pocajt, V.V., Popović, I.G., Perić-Grujić, A.A., Ristić, M.D.J., 2013. Forecasting of greenhouse gas emissions in Serbia using artificial neural networks. Energy Source. Part A 35, 733–740.
Rentziou, A., Gkritza, K., Souleyrette, R.R., 2012. VMT, energy consumption, and GHG emissions forecasting for passenger transportation. Trans. Res. A – Polym. 46, 487–500.
Rypdal, K., Winiwarter, W., 2001. Uncertainties in greenhouse gas emission inventories – evaluation, comparability and implications. Environ. Sci. Policy 4, 107–116.
Specht, D.F., 1991. A general regression neural network. IEEE Trans. Neural Netw. 2, 568–576.
Syri, S., Lehtilä, A., Ekholm, T., Savolainen, I., Holttinen, H., Peltola, E., 2008. Global energy and emissions scenarios for effective climate change mitigation – deterministic and stochastic scenarios with the TIAM model. Int. J. Greenh. Gas Control 2, 274–285.
Sözen, A., Gülseven, Z., Arcaklioğlu, E., 2007. Forecasting based on sectoral energy consumption of GHGs in Turkey and mitigation policies. Energy Policy 35, 6491–6505.
Sözen, A., Gülseven, Z., Arcaklioğlu, E., 2009. Estimation of GHG emissions in Turkey using energy and economic indicators. Energy Source. Part A 31, 1141–1159.
Thwin, M.M.T., Quah, T.-S., 2005. Application of neural networks for software quality prediction using object-oriented metrics. J. Syst. Softw. 76, 147–156.
Tripathy, M., 2010. Power transformer differential protection using neural network principal component analysis and radial basis function neural network. Simul. Model. Pract. Theory 18, 600–611.
United Nations, 1992. United Nations Framework Convention on Climate Change. In: Proceedings of the Convention on Climate Change on the Work of the Second Part of its Fifth session, NY, USA.
United Nations Economic Commission for Europe (UNEPE), 2013. http://www.unece.org
Villalba, G., Gemechu, E.D., 2011. Estimating GHG emissions of marine ports – the case of Barcelona. Energy Policy 39, 1363–1368.
Walczak, S., Cerpa, N., 1999. Heuristic principles for the design of artificial neural networks. Info. Softw. Technol. 41, 107–117.
Wang, G., Chen, S., 2012. A review on parameterization and uncertainty in modeling greenhouse gas emissions from soil. Geoderma 170, 206–216.
Ward systems group, Inc., 1993. Neuroshell 2, MD, USA.
Winiwarter, W., Rypdal, K., 2001. Assessing the uncertainty associated with national greenhouse gas emission inventories: a case study for Austria. Atmos. Environ. 35, 5425–5440.
Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. 14, 35–62.
Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.