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
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):

$$E_{i,p} = \sum_k \sum_m A_{i,k} ef_{i,k,m,p} x_{i,k,m,p} \tag{1}$$

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).

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

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).