Applying Artificial Neural Networks to Short-Term $PM_{2.5}$ Forecasting Modeling

Mihaela Oprea(✉), Sanda Florentina Mihalache, and Marian Popescu

Automatic Control, Computers and Electronics Department,
Petroleum-Gas University of Ploiesti, Ploiesti, Romania
{mihaela, sfrancu, mpopescu}@upg-ploiesti.ro

Abstract. Air pollution with suspended particles from $PM_{2.5}$ fraction represents an important factor to increasing atmospheric pollution degree in urban areas, with a significant potential effect on the health of vulnerable people such as children and elderly. $PM_{2.5}$ air pollutant concentration continuous monitoring represents an efficient solution for the environment management if it is implemented as a real time forecasting system which can detect the $PM_{2.5}$ air pollution trends and provide early warning or alerting to persons whose health might be affected by $PM_{2.5}$ air pollution episodes. The forecasting methods for PM concentration use mainly statistical and artificial intelligence-based models. This paper presents a model based protocol, *MBP – $PM_{2.5}$ forecasting* protocol, for the selection of the best ANN model and a case study with two artificial neural network (ANN) models for real time short-term $PM_{2.5}$ forecasting.

Keywords: Artificial neural networks · Forecasting modeling · Air pollution · $PM_{2.5}$ air pollutant short-term forecasting · Model based forecasting protocol

1 Introduction

Climate change is a modern topic nowadays. Air pollution is one of the most important environmental problems on the globe, and causes many types of allergies, respiratory illnesses, cardiovascular diseases, acute bronchitis diseases, etc. [1, 2]. Particulate matter (PM) is an air pollutant with high impact on humans because short-term and long-term exposure to high concentrations may produce severe health effects and premature mortality [3, 4].

Short-term forecasting of $PM_{2.5}$ air pollution trends can use different methods: deterministic, statistical, neural, hybrid (e.g. neuro-fuzzy) etc. The statistical models include linear regression, ARIMA, principal components analysis, etc., and have been used for their forecasting skills [5, 6]. The forecasted results generated using these linear statistical models are in general not satisfactory. An alternative is the use of computational intelligence approaches, such as artificial intelligence-based models [5, 7]. Artificial neural networks [8] and adaptive neuro-fuzzy inference systems (ANFIS) have been successfully applied in air pollution forecasting domain [9–11]. The chosen of an efficient forecasting method is done by experiment, depending on the available time series databases with measurements of $PM_{2.5}$ concentration, meteorological parameters, other air pollutants concentration that influence $PM_{2.5}$. Depending on the

L. Iliadis and I. Maglogiannis (Eds.): AIAI 2016, IFIP AICT 475, pp. 204–211, 2016.
DOI: 10.1007/978-3-319-44944-9_18

correlation degree with $PM_{2.5}$, a part of these parameters can be considered as inputs in the $PM_{2.5}$ forecasting model. We are applying such a model under the ROKIDAIR research project (http://www.rokidair.ro) whose goal is to provide an intelligent tool (ROKIDAIR DSS) for early warning/alerting of $PM_{2.5}$ air pollution episodes in urban areas (in two pilot cities from Romania, Ploiesti and Targoviste), in order to reduce the potential negative effects of air pollution on children health. Within this project we are developing a model based on artificial intelligence, named ROKIDAIR IA which has two main components: a short-term $PM_{2.5}$ forecasting component and an intelligent decision support component, based on knowledge. In this paper we focus on short-term $PM_{2.5}$ forecasting modeling based on ANN.

2 The Artificial Neural Network Approach for Short-Term $PM_{2.5}$ Forecasting

Artificial neural networks are universal approximators that can learn complex mapping between the input and the output data [12]. An ANN is composed by a set of artificial neurons which are connected according to a topology. Each connection between two neurons has a weight (a numerical value in the interval [0, 1]) showing the degree of that connection which is derived during the ANN training stage. The number of input neurons is given by the input parameters of the forecasting problem, the output neurons are the $PM_{2.5}$ forecasted values in the time window t + k (named also, forecast horizon), while the number of hidden neurons is derived by experiment during training. Some of the ANNs types most used to solve forecasting problems are feed forward artificial neural networks [13], recurrent ANNs [14] and radial basis ANNs [12]. Some recent research results reported in the literature confirmed the good performance of the neural predictors used to detect the air pollution evolution [15–17].

Figure 1 shows an example of a feed forward ANN for $PM_{2.5}$ forecasting. The model uses past measurements of $PM_{2.5}$ concentration and other atmospheric parameters. The ANN has an input layer, an output layer and one or more hidden layers. Usually, one hidden layer is enough to capture the evolution of the forecasted parameter according to the data sets available for ANN training. Feed forward ANNs are trained with a backpropagation algorithm which can be improved by choosing the right learning parameters, adjusted during training. The generation of an ANN model must follow three steps: (1) ANN training with a training algorithm on a training data set; (2) ANN validation on a training data set; (3) ANN testing on a testing data set.

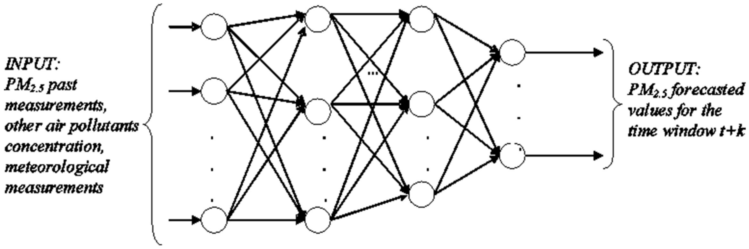
INPUT:
PM2.5 past measurements, other air pollutants concentration, meteorological measurements
OUTPUT:
PM2.5 forecasted values for the time window t+k

Fig. 1. Example of a feed forward ANN for $PM_{2.5}$ forecasting

The $PM_{2.5}$ ANN forecasting model is derived by training the ANN on a training set selected from the data sets that are available for the urban area that is studied. After training the ANN model is validated and tested on specific data sets. A recent comparison between some ANN models applied to $PM_{2.5}$ prediction is described in [18]. The main advantage of an ANN forecasting model is given by its capability to capture with good accuracy the forecasting function when enough large data sets are used. Our proposed approach for $PM_{2.5}$ short-term forecasting is based on the *MBP - $PM_{2.5}$ forecasting* protocol, developed under the ROKIDAIR project.

3 The $PM_{2.5}$ Forecasting Model Development Protocol

We have developed a protocol, *MBP - $PM_{2.5}$ forecasting*, for building the $PM_{2.5}$ forecasting model under the ROKIDAIR project. The main purpose of the protocol is to facilitate the systematic construction of the short-term $PM_{2.5}$ forecasting model that will be used by the ROKIDAIR Decision Support System in order to provide decisions under the form of warning/alerting messages regarding the potential negative effects on children health of the $PM_{2.5}$ air pollution episodes. The *MBP - $PM_{2.5}$ forecasting* protocol defines the steps of $PM_{2.5}$ forecasting model design. The air pollution forecasting module determines the evolution for short term $PM_{2.5}$ concentration.

Figure 2 presents the logic diagram of the *MBP - $PM_{2.5}$ forecasting* protocol (with 4 main steps) for the short-term $PM_{2.5}$ forecasting module of the ROKIDAIR Decision Support System.

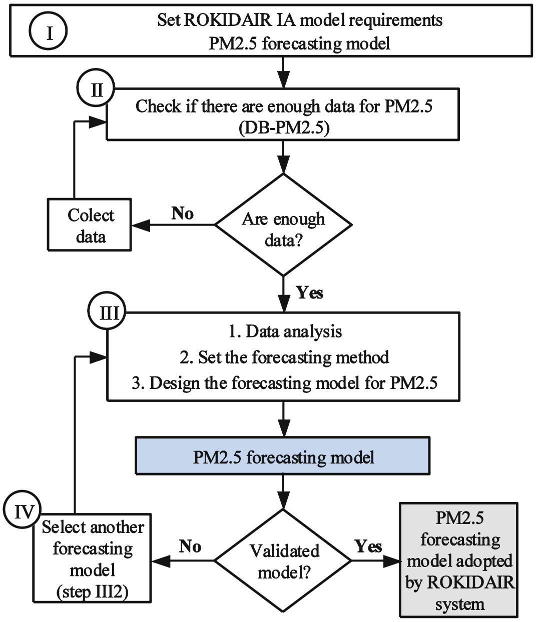
I
Set ROKIDAIR IA model requirements
PM2.5 forecasting model
II
Check if there are enough data for PM2.5
(DB-PM2.5)
Colect data
No
Are enough data?
Yes
III
1. Data analysis
2. Set the forecasting method
3. Design the forecasting model for PM2.5
PM2.5 forecasting model
IV
Select another forecasting model (step III2)
No
Validated model?
Yes
PM2.5 forecasting model adopted by ROKIDAIR system

Fig. 2. Logic diagram of the *MBP-$PM_{2.5}$ forecasting* protocol (steps I, II, III, IV)

In the first step are set the $PM_{2.5}$ forecasting model requirements (as e.g. past measurements window time, forecasting horizon, input/output parameters, forecasting accuracy). In step II it is checked if the database with $PM_{2.5}$ concentration measurements and other $PM_{2.5}$ related atmospheric parameters measurements (DB-$PM_{2.5}$) has enough data. If there are not enough data, it is started a process of data collection (usually, for the analyzed urban area or a similar $PM_{2.5}$ air polluted urban area). When enough data are stored in the database, step III is performed with a data analysis sub-step (III.1), followed by the setting of the forecasting method (III.2) and the design of the $PM_{2.5}$ forecasting model (III.3) according to the methodology of selecting the best solution. After step III, the short-term $PM_{2.5}$ forecasting model is generated. If the model is not validated, another forecasting method is chosen in step III.2. If the model is validated than it is adopted by the ROKIDAIR system. The model validation is performed according to the desired forecasting performance which is measured with some indicators: mean absolute error (MAE), index of agreement (IA), root mean square error (RMSE), and coefficient of determination (R^2).

As we are focusing on the artificial neural network based forecasting method, we present the main steps of the methodology proposed for feed forward and radial basis ANN model selection which were integrated in the ROKIDAIR *MBP – $PM_{2.5}$ forecasting* protocol.

MBP – $PM_{2.5}$ forecasting protocol – ANN Selection Methodology

Step 1. Time series data processing (i.e. DB-$PM_{2.5}$) – in order to be used by the $PM_{2.5}$ ANN forecasting method;
Step 2. Select the most relevant atmospheric input parameters to short-term $PM_{2.5}$ forecasting (e.g. by using principle component analysis);
Step 3. Select the training, validation and testing data sets for the ANN model;
Step 4. Set the ANN architecture (e.g. input nodes, output nodes, hidden nodes, radial function, cluster seed, number of clusters etc.);
Step 5. Adjust the training parameters according to the training algorithm;
Step 6. ANN training, validation, testing using the training data set chosen in step 3;
Step 7. Analyze the performances of the designed ANN model (i.e. RMSE, IA, R^2);
Step 8. Select the best ANN model for real time short-term $PM_{2.5}$ forecasting.

A good $PM_{2.5}$ forecasting model should have a smaller error (RMSE, MAE), a coefficient of determination and an index of agreement close to 1. In order to keep the $PM_{2.5}$ short-term forecasting model as simple as possible for an efficient real time $PM_{2.5}$ forecasting, a minimum number of the atmospheric parameters (e.g. temperature and relative humidity) most relevant to $PM_{2.5}$ concentration evolution are chosen.

4 Experimental Results

The data sets used in this study come from an air quality monitoring station from an urban area of Ploiesti, Romania, and each data set contains approximately 4200 samples for $PM_{2.5}$ concentrations and temperature. From all meteorological parameters the temperature is correlated with $PM_{2.5}$ evolution. The data from Ploiesti monitoring station referring to $PM_{2.5}$ concentrations has the maximum of 36.45 $\mu g/m^3$, and a minimum of 0.19 µg/m^3. In the same time, the temperature data set has the maximum of 37.24 °C, and the minimum of −0.2 °C.

The proposed forecasting models use normalized data for both $PM_{2.5}$ concentrations and temperature. The data were randomly divided with the following percentages: 70 % for training, 15 % for validation and 15 % for testing. We propose two types of forecasting models in this study, based on ANNs. One model has as inputs the four previous $PM_{2.5}$ hourly concentrations (Fig. 3a) and the other has one more input than the first one, namely the current hourly temperature (Fig. 3b). The output of the models is the same in both cases - short term forecasted value for the next hour $PM_{2.5}$ concentration.

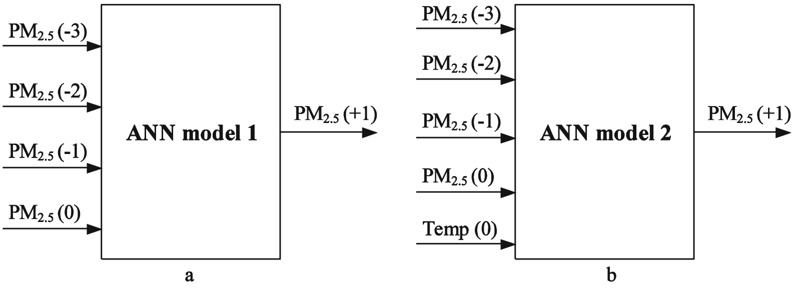

Fig. 3. ANN models

The structure of the proposed neural network contains four neurons in the input layer, one hidden layer and one neuron in the output layer. In the study there were used two types of neural networks, namely feed forward backpropagation (FFwd) and layer recurrent (LRec). As training algorithm the preferred method is Levenberg-Marquardt, and for the adaptive learning functions there are studied the gradient descent with momentum weight and bias (learngdm) and gradient descent weight and bias (learngd). The simulations were performed modifying also the number of neurons in the hidden layer.

The training and validation errors have values around 0.001 and 0.0007 respectively. The accuracy of the models can be evaluated based on the comparison between the actual value and forecasted value of $PM_{2.5}$ concentration, with mean error and standard deviation criteria. The performances of the designed ANN models are compared using statistical indices such as RMSE, IA, R^2, and R.

The two models are compared using statistical criteria and a selection of the results are presented in Tables 1 and 2, the best configuration for each ANN model being highlighted.

For the first model using only PM concentrations as inputs the best results are obtained in the case of layer recurrent structure with 5 neurons in the hidden layer and the *learngdm* adaptation learning function. In this case the root mean squared error have the smallest value, and IA, R^2 and R indices have the biggest values.

Table 1. Statistical indices for ANN model 1

ANN model 1		RMSE [μg/m^3]	IA	R^2	R
4 × 3 × 1/ Learngdm	FFwd	1.1106	0.9902	0.9620	0.9809
	LRec	1.1268	0.9899	0.9609	0.9803
4 × 3 × 1/ Learngd	FFwd	1.1123	0.9902	0.9619	0.9808
	LRec	1.1132	0.9902	0.9619	0.9808
4 × 4 × 1/ Learngdm	FFwd	1.1188	0.9901	0.9615	0.9806
	LRec	1.1257	0.9899	0.9610	0.9803
4 × 4 × 1/ Learngd	FFwd	1.1188	0.9901	0.9615	0.9806
	LRec	1.1182	0.9900	0.9615	0.9806
4 × 5 × 1/ Learngdm	FFwd	1.1128	0.9902	0.9619	0.9808
	LRec	**1.0908**	**0.9905**	**0.9634**	**0.9815**
4 × 5 × 1/ Learngd	FFwd	1.1132	0.9901	0.9618	0.9808
	LRec	1.1193	0.9901	0.9614	0.9806
4 × 6 × 1/ Learngdm	FFwd	1.1057	0.9903	0.9624	0.9810
	LRec	1.1110	0.9902	0.9620	0.9809
4 × 6 × 1/ Learngd	FFwd	1.1064	0.9903	0.9623	0.9810
	LRec	1.0933	0.9905	0.9632	0.9814

Table 2. Statistical indices for ANN model 2

ANN model 2		RMSE [μg/m^3]	IA	R^2	R
4 × 3 × 1/ Learngdm	FFwd	1.1106	0.9902	0.9620	0.9809
	LRec	1.1110	0.9902	0.9620	0.9808
4 × 3 × 1/ Learngd	FFwd	1.1123	0.9902	0.9619	0.9808
	LRec	1.1198	0.9900	0.9614	0.9806
4 × 4 × 1/ Learngdm	FFwd	1.0985	0.9904	0.9629	0.9813
	LRec	1.1206	0.9900	0.9613	0.9805
4 × 4 × 1/ Learngd	FFwd	1.1188	0.9901	0.9615	0.9806
	LRec	1.1046	0.9903	0.9624	0.9811
4 × 5 × 1/ Learngdm	FFwd	1.0966	0.9905	0.9630	0.9813
	LRec	1.1134	0.9901	0.9618	0.9808
4 × 5 × 1/ Learngd	FFwd	1.1221	0.9900	0.9612	0.9805
	LRec	1.0998	0.9904	0.9628	0.9812
4 × 6 × 1/ Learngdm	FFwd	1.1074	0.9902	0.9622	0.9810
	LRec	1.1256	0.9899	0.9610	0.9803
4 × 6 × 1/ Learngd	**FFwd**	**1.0951**	**0.9905**	**0.9631**	**0.9814**
	LRec	1.0966	0.9904	0.9630	0.9814

The second model with temperature as additional input has the best results (comparing the same statistical indices) in the case of feed forward structure with 6 neurons in the hidden layer and the *learngd* adaptation learning function.

The best results from the two models showed that no significant enhancement has been produced when current hourly temperature is included as additional input variable to the second ANN model. The best structure between the two is the one from the first model with PM concentrations as inputs (4 × 5 × 1 – Learngdm – Layer Recurrent) with: RMSE = 1.0908 μg/m^3, IA = 0.9905, R^2 = 0.9634 and R = 0.9815.

Figure 4 presents a partial view of the comparison between testing and forecasted data for the best ANN structure.

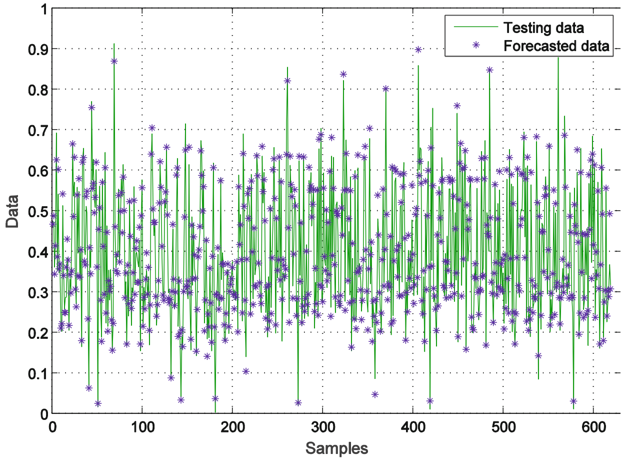

Fig. 4. Comparison between testing and forecasted data

5 Conclusions

The paper presented two ANN models proposed for real time $PM_{2.5}$ short-term forecasting in the case of a polluted town in Romania. In order to select the best ANN forecasting model, we have designed a model based $PM_{2.5}$ forecasting protocol, named *MBP – $PM_{2.5}$ forecasting* protocol, which is integrated in the ROKIDAIR MBP protocol for the development of the ROKIDAIR DSS. The first proposed model uses as inputs only hourly PM concentrations and the second one uses an additional input the current hourly temperature. The conclusions are that the accuracy of both ANN models are almost the same, so both models can be considered appropriate approaches to real time short term forecast. As future work we propose to include other meteorological variables into the model, use additional hybrid modelling techniques such as FIR with genetic algorithm, or expand the forecasting window to next day.

Acknowledgements. The research leading to these results has received funding from EEA Financial Mechanism 2009-2014 under the project ROKIDAIR "*Towards a better protection of children against air pollution threats in the urban areas of Romania*" contract no. 20SEE/30.06.2014.

References

1. Kampa, M., Castanas, E.: Human health effects of air pollution. Environ. Pol. **151**, 362–367 (2008)
2. Qin, G., Meng, Z.: Effects of sulfur dioxide derivatives on expression of oncogenes and tumor suppressor genes in human bronchial epithelial cells. Food Chem. Toxicol. **47**, 734–744 (2009)
3. Baker, K.R., Foley, K.M.: A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM2.5. Atmos. Environ. **45**, 3758–3767 (2011)
4. Nebot, A., Mugica, F.: Small-particle pollution modeling using fuzzy approaches. In: Obaidat, M.S., Filipe, J., Kacprzyk, J., Pina, N. (eds.) Simulation and Modeling Methodologies. AISC, vol. 256, pp. 239–252. Springer, Heidelberg (2014)
5. Oprea, M., Dragomir, E.G., Mihalache, S.F., Popescu, M.: Prediction methods and techniques for PM2.5 concentration in urban environment (in Romanian). In: Iordache, S., Dunea, D. (eds.) Methods to Assess the Effects of Air Pollution with Particulate Matter on Children's Health (in Romanian), pp. 387–428. MatrixRom, Bucharest (2014)
6. Kumar, N., Chu, A., Foster, A.: An empirical relationship between PM2.5 and aerosol optical depth in Delhi metropolitan. Atmos. Environ. **41**(21), 4492–4503 (2007)
7. Akkoyunlu, A., Yetilmezsoy, K., Erturk, F., Oztemel, E.: A neural network-based approach for the prediction of urban SO2 concentrations in the Istanbul metropolitan area. Inter. J. Environ. Pol. **40**, 301–321 (2010)
8. Yilmaz, I., Kaynar, O.: Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Syst. Appl. **38**, 5958–5966 (2011)
9. Morabito, F.C., Versaci, M.: Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data. Neural Netw. **16**, 493–506 (2003)
10. Yildirim, Y., Bayramoglu, M.: Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere **63**, 1575–1582 (2006)
11. Ashish, M., Rashmi, B.: Prediction of daily air pollution using wavelet decomposition and adaptive-network-based fuzzy inference system. Int. J. Environ. Sci. **2**(1), 185–196 (2011)
12. Haykin, S.: Neural networks. A comprehensive foundation. Pearson Education Inc., New Delhi (1999)
13. Hornik, K.: Approximation capabilities of multilayer feed-forward networks. Neural Netw. **4**, 251–257 (1991)
14. Mandic, D., Chambers, J.: Recurrent Neural Networks for Prediction Learning Algorithms, Architectures and Stability. Wiley, New York (2001)
15. Kurt, A., Oktay, A.B.: Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl. **37**, 7986–7992 (2010)
16. Fernando, H.J., Mammarella, M.C., Grandoni, G., Fedele, P., Di Marco, R., Dimitrova, R., Hyde, P.: Forecasting PM10 in metropolitan areas. Efficacy of neural networks. Environ. Pollut. **163**, 62–67 (2012)
17. Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J.: Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. **107**, 118–128 (2015)
18. Oprea, M., Mihalache, S.F., Popescu M.: A comparative study of computational intelligence techniques applied to $PM_{2.5}$ air pollution forecasting. In: Proceedings of 2016 6th International Conference on Computers Communications and Control (ICCCC), pp. 103–108. Baile Felix, Oradea, Romania (2016)