Data Mining and ANFIS Application to Particulate Matter Air Pollutant Prediction. A Comparative Study

Mihaela Oprea[1], Marian Popescu[1], Sanda Florentina Mihalache[1] and Elia Georgiana Dragomir[2]
[1]*Automatic Control, Computers and Electronics Department, Petroleum-Gas University of Ploiesti, Ploiesti, Romania*
[2]*Informatics, Information Technology, Mathematics and Physics Department, Petroleum-Gas University of Ploiesti, Bd. Bucuresti, 39, Ploiesti, Romania*

Keywords: Prediction Model, Data Mining, Adaptive Neuro-Fuzzy Inference System, Particulate Matter Air Pollution.

Abstract: The paper analyzes two artificial intelligence methods for particulate matter air pollutant prediction, namely data mining and adaptive neuro-fuzzy inference system (ANFIS). Both methods provide predictive knowledge under the form of rule base, the first method, data mining, as an explicit rule base, and ANFIS as an internal fuzzy rule base used to perform predictions. In order to determine the optimal number of prediction model inputs, we have perform a correlation analysis between particulate matter and other air pollutants. This operation imposed NO_2 and CO concentrations as inputs of the prediction model, together with four values of PM_{10} concentration (from current hour to three hours ago), the output of the model being the prediction of the next hour PM_{10} concentration. The two prediction models are investigated through simulation in different structures and configurations using SAS® and MATLAB® respectively. The results are compared in terms of statistical parameters (RMSE, MAPE) and simulation time.

1 INTRODUCTION

Artificial intelligence (AI) provides very good prediction models for a variety of applications in domains such as engineering, environmental science, meteorology, economy, medicine, finance, banking, education etc. They find in shorter time, sub-optimal solutions, being proper for use in real time systems.

Developing more accurate real time air pollution forecasting systems is nowadays an important interdisciplinary research topic for several academic communities jointly working in environmental science (air pollution), meteorology, computer science, artificial intelligence, statistics, physics etc. One of the harmful air pollutant in urban areas which can cause significant health problems especially to sensitive people (such as children, elderly) is particulate matter (PM). As smaller is the PM diameter size as much significant is the potential negative effect on human health. PM_{10} and $PM_{2.5}$ are two types of PM that need careful analysis of their concentration levels and accurate prediction for short-terms (as e.g. next hours, next day) in order to reduce the impact on human health when higher levels are recorded. Early warning systems based on accurate prediction of air pollution can help to improve life quality in urban areas.

The research topic tackled in this paper focus on the comparison of two artificial intelligence prediction models, data mining (DM) and adaptive-network fuzzy inference system (ANFIS) that have the potential to increase the prediction accuracy for PM air pollutant. The selection of the two methods was done taking into account the recent research results reported in the literature (primarily the atmospheric environment science and artificial intelligence literature) and previous research work results reported in (Oprea et. al., 2016b). We made a comparison between them in order to identify the best one which can be used in the Ploiesti city from Romania, a city that has a higher air pollution.

The purpose of our research work is to develop a real-time PM air pollutant forecasting system which provides next hours PM concentration level with a higher accuracy.

The paper is organized as follows. Section 2 presents an overview on prediction models. The two artificial intelligence prediction models, data mining and ANFIS, are described in section 3. The experimental results are detailed and discussed in section 4. The final section concludes the paper and identifies some future work.

Oprea M., Popescu M., Mihalache S. and Dragomir E.
Data Mining and ANFIS Application to Particulate Matter Air Pollutant Prediction. A Comparative Study.
DOI: 10.5220/0006196405510558
In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017)*, pages 551-558
ISBN: 978-989-758-220-2

2 AN OVERVIEW ON PREDICTION MODELS

The main types of prediction models that are currently used in forecasting environmental (air, water, soil) pollution are: climatology models, deterministic models, statistical models, artificial intelligence models, hybrid models.

Computational intelligence (CI) models are the most used artificial intelligence prediction models. They include fuzzy inference systems (FIS), artificial neural networks (ANNs), genetic algorithms (GAs), swarm intelligence models (such as ant colony optimization - ACO, particle swarm optimization - PSO, bees colony - ABC etc) as well as combinations of them (e.g. ANFIS which combines FIS and ANN). Several applications were reported for time series forecasting using CI models (see e.g. (Palit and Popovic, 2005)).

The prediction models used for real time air quality forecasting are (Zhang et al., 2012): simple empirical approaches (e.g. climatology), parametric (including statistical) models (e.g. ANNs, regression trees, fuzzy models), advanced physically-based approach (e.g. deterministic – CTM models).

Three prediction models were used for $PM_{2.5}$ forecasting in (Perez and Salini, 2008): a linear model, a multilayer neural network and a hybrid clustering algorithm. All methods proved to be good, but the last one gave more accurate results in detecting $PM_{2.5}$ high level concentrations. Another research work (Elangasinghe et al., 2014) combines ANN with k-means clustering for a better understanding of PM_{10} and $PM_{2.5}$ complex time series measured in a coastal site of New Zealand. In this case the inclusion of clustering rankings in the ANN model improved the prediction accuracy when PM higher concentrations are registered.

Air quality prediction using fuzzy logic and statistical models (autoregressive) is discussed in (Carbajal-Hernández et al., 2012). Another research which apply fuzzy logic to time series forecasting is presented in (Domańska and Wojtylak, 2012).

A case study of ANFIS modelling for air pollution in Serbia is described in (Savić et al., 2014). Another research work using ANFIS combined with a data preprocessing technique (output-dependent data scaling) to predict daily levels of PM_{10} in Konya city, Turkey, is detailed in (Polat et al., 2012). Some recent research work that report more accurately prediction results by using ANFIS prediction models in environmental systems are introduced in (Mekanik et al., 2015) - for seasonal rainfall forecasting in southeast Australia, (Hajek and Olej, 2015) - for common air quality index prediction in some regions from Czech Republic, (Mishra et al., 2015) – for $PM_{2.5}$ forecast during haze episodes in Delhi, India.

Shahraiyni et al. (2015) proposes an identification scheme for selecting meaningful inputs to ANFIS model used to simulate the virtual air pollution monitoring stations from Berlin. The hourly data sets for particulate matter PM_{10} are converted into daily mean for the available data. The results show a reduced computing time for inferring a smaller number of ANFIS rules. The proposed ANFIS models have good results in terms of statistical indices. Ausati and Amanollahi (2016) studies ANFIS method and statistical methods for daily average $PM_{2.5}$ prediction in a polluted urban area of Iran. The hybrid models offer the best solutions according to statistical indices.

Prasad and al. (2016) want to predict five air pollutants daily concentration (PM_{10} among them) based on ANFIS using as inputs several meteorological parameters and previous' day air pollutant concentration. The datasets is for an urban region from India. Forward selection method is used to reduce the number of ANFIS inputs, reducing the computational time and effort.

A hybrid method, statistical and ANFIS combined is proposed to predict the daily PM_{10} concentration from an urban area from Turkey in Polat (2012). The proposed ANFIS use four meteorological parameters as additional inputs to the previous PM_{10} daily concentration.

Oprea et al. (2016a) presents a comparative study for $PM_{2.5}$ prediction between ANN and ANFIS. The used dataset contains data from an urban region from Germany with hourly concentrations. The inputs of ANFIS are based only on previous hourly PM concentration. Mihalache et al. (2015) test the ANFIS prediction method on three data sets from different urban region from Romania. The next hour PM_{10} concentration is the ANFIS output and the inputs are based only on previous hourly concentrations.

A combination of FIS and GA for air pollution prediction based on GIS data is presented in (Shad et al., 2009).

Data mining models used for prediction are reported in (Osrodka et al., 2005) – for high level air pollution forecasting in urban industrial area from southern Poland, (Siwek and Ossowski, 2016) – which use GA and a linear method for feature selection and random forest (forming an ensemble of decision trees) and ANNs as prediction models. Another research work on data mining models for air quality prediction in Athens, Greece, is described in (Riga et al., 2009).

The research results reported so far highlighted the potential benefits of using data mining techniques as predictive models for air pollutants concentrations.

A combination between ANN and inductive learning algorithms applied to prediction is discussed in (Bae and Kim, 2011). Deep recurrent ANN is another promising method for $PM_{2.5}$ prediction (see e.g. a recent research work reported in (Ong et al., 2016). The application of a predictor's ensemble for daily average PM_{10} forecasting is tackled in (Siwek et al., 2011). The use of decision trees and neural networks proved to give also, good prediction accuracy of air quality forecast (see e.g. (Loya et al., 2013)).

The main conclusion of our overview is that most of the prediction models based on artificial intelligence techniques are chosen as a function of the air quality monitoring area specific data, and the hybrid models built as a combination of several techniques, gave the best results. However, some of the AI techniques, such as data mining, ANFIS and ANNs need deeper investigation of their potential in providing more accurate forecasts. Moreover, few research papers presents comparative studies between data mining techniques (such as decision tree) and ANFIS. Our research work focus on the analysis of a data mining technique - decision tree (CHAID algorithm) and ANFIS (Takagi-Sugeno) in solving short-term prediction of PM_{10} air pollutant in the city of Ploiesti, Romania. Several experiments were performed and the main lessons that were learnt are presented in this paper.

3 METHODS DESCRIPTION

A brief description of the two methods that were selected in our research study, data mining – decision trees and ANFIS is given as follows.

3.1 Data Mining – Decision Tree

Data mining can be defined as the process of knowledge (pattern) extraction from large databases. It comprises a set of techniques from statistics (e.g. statistical classifiers) and artificial intelligence (e.g. computational intelligence techniques and rule-based systems).

Some examples of data mining techniques are: decision tree classifiers (e.g. C4.5, CHAID, REP Tree, Random Forest), support vector machine (SVM), instance-based classifiers, artificial neural networks (such as multi-layer perceptron - MLP and radial basis function neural network – RBF-NN), rule-based classifiers.

We have chosen to analyze decision trees as they have important characteristics: offer fast and computationally inexpensive prediction; they are nonparametric; provide an intuitive representation of extracted knowledge, being very useful for environmental data; they generate a rule base.

The decision tree method is often used for gaining information in a decision-making process.

A decision tree is a tree-like structure where each branch is a possible choice and each leaf node is a decision. This technique classifies instances by crossing the tree from the root node to leaf nodes. It starts by testing the root attribute, then by moving the tree branches according to data attribute values in the given data set. Attributes in a classification problem are usually of two types: nominal or numerical (their values are real numbers).

At each step there is selected a variable which is considered "best". Different type of decision trees use different formula for this. One of the possibility is the information gain (given by (1)).

$$IG(M,A) = En(M) - \sum_{v \in Values(A)} \frac{|M_v|}{M} En(M_v) \qquad (1)$$

Parameter M is a training data set, A - an attribute, $En()$ - the function which calculates the entropy, $Values(\mathrm{v})$ - a data set with all possible values of A and M_v represents a subset of M in which A has value v.

Decision trees generate a set of rules that may be useful in predicting a new set of data. This knowledge can be used in anticipation of a possible air pollution episode.

3.2 ANFIS Prediction Method

ANFIS technique can be used as a prediction method. The available dataset impose the inputs used to predict a specific variable that becomes the ANFIS output. The prediction is usually on short term (next hour, next two hours, next 6 hours). The inputs must be relevant to the output variable due to the increase of computational effort with additional inputs. For example an ANFIS with 4 inputs each described by 3 membership functions generates 81 rules. Adding one input increases the rules number to 243 rules. A statistical method can be used to select between relevant inputs to the selected output prediction. Then the input-output dataset is randomly divided into training, validating and testing datasets. The starting FIS can have an empty rule base or an existing number of fuzzy rules. In the first case ANFIS must generate the rules with respect to continuity, consistency and completeness required for fuzzy rule base. This is the case where there are no obvious *if-then* rules between the inputs and outputs. In the second case, from experience one can generate a fixed number of fuzzy rules and then the ANFIS has to adjust the premise parameters and consequent parameters to match the input output datasets. For example, if there are two rules:

$$\textit{If } U_1 \textit{ is } \mu_{A1} \textit{ and } U_2 \textit{ is } \mu_{B1} \textit{ then } f_1 = c_1U_1 + d_1U_2 + r_1 \quad (2)$$

$$\textit{If } U_1 \textit{ is } \mu_{A2} \textit{ and } U_2 \textit{ is } \mu_{B2} \textit{ then } f_2 = c_2U_1 + d_2U_2 + r_2 \quad (3)$$

where μ_{Ai} and μ_{Bi} are input membership functions and c_i, d_i, and r_i are the constant values for Takagi Sugeno part, the ANN part adjust the values of the consequent parameters (c_i, d_i, and r_i) and the parameters associated with premise parameters (membership functions μ_{Ai} and μ_{Bi}). ANFIS technique used as prediction for PM usually has to model complicated nonlinear dependencies between inputs and output, therefore the rule base is empty and ANFIS has to generate the rules, and adjust the FIS parameters to match the input output datasets. The resulted FIS architecture is evaluated with the testing data and the prediction is compared to the testing data via statistical indices.

The ANFIS performance can be adjust modifying the methods and parameters associated: the method of generating the FIS structure, the number of inputs, the shapes of membership functions defined for the input variables, the granularity of each input, the type of output function, the optimization method to train FIS and the number of training epochs.

4 EXPERIMENTAL RESULTS

4.1 Data Set

The data used in this study are from an air quality monitoring station from Ploiești, Romania, and contains hourly concentrations of the air pollutants PM_{10}, SO_2, CO, NO_2, and NO_x, each pollutant having around 6000 samples.

In order to determine which pollutants have greater influence on PM_{10} concentration the correlation coefficient is calculated its values being presented in Table 1.

From Table 1 it can be observed that CO and NO_2 provide the highest values of the correlation coefficient. Thus, these pollutants will be used together with PM_{10} as inputs in the prediction models from this study.

Table 1: Correlation Coefficient.

Pollutant	Cor. Coef.
SO_2	0.048
CO	0.603
NO_2	0.556
NO_x	0.480

The data set consisting of PM_{10}, CO and NO_2 concentrations has the following characteristics:

- PM_{10} data have a maximum of 261.33 µg/m^3, and an average of 45.05 µg/m^3;
- The maximum CO concentration is 1.8 mg/m^3, with an average of 0.33 mg/m^3;
- NO_2 concentration has a maximum of 79.91 µg/m^3, and an average of 27.3 µg/m^3.

All experiments presented in the following are based on a data set divided into 70% for training, 15% for validation and 15% for testing.

4.2 Decision Tree Method

The software package used in the decision tree experiments includes SAS® Enterprise Guide® and SAS® Enterprise Miner™. SAS implementation involves the use of a multi-way decision trees (Schubert and Lee, 2011). It can be chosen the splitting criteria and other options that determine the method of tree construction. The options include the popular features of CHAID (Chi-square automatic interaction detection) and those described in (Breiman et al., 1984). The evaluation criteria of selection rules can be based either on the results of tests to determine certain statistical parameters (such as the F-test and Chi-Square, methods that accept an input value p as stop criterion) or on reducing variance, entropy or Gini parameter. The F test and variance can be used for interval values.

The decision tree experiments test different values for the target variable criteria analysis (TVCA), missing values (MV), branching factor (BF) and tree depth (TD).

The two methods that have been tested to assess possible variables used in splitting rules are ProbF (the p-value of the F test that is associated with the node variance) and Variance (the reduction in the square error from the node means).

Decision trees implemented in SAS Enterprise Miner can handle missing values in three ways: *Use in search* (UIS - using the missing values when calculating worth a rule of splitting; this will always produce a splitting rule that assigns the missing values to the branch that maximizes the worth of the split), *Largest Branch* (LB - assigns the observations that contain missing values to the largest branch) and *Most Correlated Branch* (MCB - missing values are assigned to the branch with the lowest residual sum of squares calculated) (SAS Ent. Miner, 2016).

The representative tests to determine the most appropriate process method for missing values and target variable are summarized at the beginning of Table 2. The branching and depth factor of decision tree were kept at node default settings. Thus, the smaller value for the mean square error in the validation set (V_ASE) is 137.17 obtained by the DT2 tree using *Variance* and *Use in search* methods.

Table 2: The most relevant decision tree experiments.

Tree	TVCA	MV	BF	TD	V_ASE
DT1	ProbF	UIS	2	6	156.50
DT2	Variance	UIS	2	6	137.17
DT3	ProbF	LB	2	6	157.74
DT4	ProbF	MCB	2	6	156.50
DT5	Variance	LB	2	6	138.42
DT6	Variance	MCB	2	6	137.17
DT7	Variance	UIS	3	6	149.75
DT11	Variance	UIS	7	6	140.44
DT14	Variance	UIS	2	7	136.53
DT15	Variance	UIS	2	8	135.33
DT16	Variance	UIS	3	4	153.92
DT17	Variance	UIS	3	8	149.75
DT19	Variance	UIS	5	8	162.94

The most relevant experiments made to determine optimum values for the depth factor and decision tree branching are presented in second part of Table 2. All decision trees have set *Variance* and *Use in search* methods. It is noted that DT15 has 135.33 the lowest V_ASE value. For a branching factor greater than 2 the V_ASE begins to grow again. The DT15 identifies 8 as optimal depth value.

In Figure 1 is sketched a partial view of DT15 decision tree. The root node selected PM10_t parameter, the amount of PM_{10} measured at the t moment influence the changes in PM_{10} concentration in the next hour. The same importance for this parameter is considered by the nodes from level 1, 2 and partial level 3 in the tree.

The next parameter in generating the decision tree is NO2_t, followed by CO_t. At the 5th level are being selected the other database pollutants: PM10_t_1, PM10_t_2, PM10_t_3. A possible explanation can be that the current concentrations most influence the next hour PM_{10} value compared with other time moments measured data.

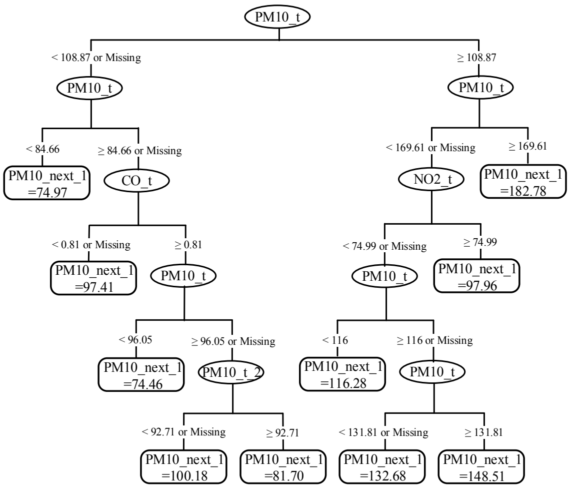

Figure 1: Partial view of DT15 decision tree.

Table 3: The model score on distributed intervals.

Range for Predicted	Mean Target	Mean Predicted	Model Score
141.572 - 148.518	145.032	148.518	145.045
113.787 - 120.733	81.26	116.288	117.26
99.895 - 106.841	105.526	100.183	103.368
92.949 - 99.895	79.573	97.549	96.422
79.056 - 86.003	87.78	81.703	82.53
72.110 - 79.056	75.53	75.025	75.583
58.218 - 65.164	61.339	59.822	61.691
51.272 - 58.218	54.016	52.757	54.745
44.326 - 51.272	41.958	46.766	47.799
37.380 - 44.326	37.027	39.175	40.853
30.433 - 37.380	31.361	32.309	33.906
23.487 - 30.433	23.977	27.473	26.96
16.541 - 23.487	21.754	21.456	20.014
9.595 - 16.541	10.493	11.246	13.068

Table 3 presents the model score on distributed intervals for the validation data set.

Statistical results of the leaf nodes are shown in Table 4. There are selected only the nodes that presented the pattern usage degree greater than 0.70. The pattern usage degree is the ratio of the number of observations in the branch to the number of observations in the root node (SAS Ent. Miner, 2016).

Table 4: Leaf nodes statistical results (selection by pattern usage degree).

Node id	Parent id	Node depth	Predicted: PM10_next_1	Validated: PM10_next_1	RASE	VRASE	Pattern usage degree
21	10	4	46.48	41.41	8.73	13.17	1.03
36	19	5	31.13	29.66	7.25	8.48	0.98
37	19	5	35.45	32.21	8.96	5.74	0.94
43	22	5	58.57	59.82	8.24	10.09	0.90
33	17	5	21.29	21.88	4.25	6.15	0.88
35	18	5	25.21	19.00	4.09	14.94	0.88
68	37	6	36.59	32.92	7.84	5.25	0.84
60	33	6	21.68	22.33	4.04	6.13	0.81
80	43	6	59.82	61.34	8.24	10.24	0.72
97	60	7	22.21	22.32	3.86	6.33	0.72

Node 21 has as parent the node 10, is situated at the 4th level depth, it is characterized by a value of 8.73 for the RASE training set, and 13.17 V_RASE value in the validation set. The PM10_next_1 predicted value is 46.48 and the PM10_next_1 validated value is 41.41. Some IF-THEN rules generated by the decision tree are shown in Table 5. Rule 1 is specific to the node 12 and predicts an extremely high value 74.97 for PM10_next_1 if PM10_t is in the range [67.35, 84.66]. Rule 2 of the node 23 identifies the value 43.53 of the target variable if PM10_t is between 49.29 and 67.35 and if NO2_t is greater than 57.59.

The knowledge extraction generated rules generally capture characteristics that influences PM_{10} concentration evolution. Thus, the forecast model proposed has identified influence of the PM_{10} concentrations measured in a previous moment of time as well as other atmospheric parameters measured locally (NO_2, CO).

Table 5: Examples of knowledge extraction rules.

Rule	Rule description
1	**IF** PM10_t < 84.66 **AND** PM10_t >= 67.35 **THEN** Predicted: PM10_next_1 = 74.97
2	**IF** PM10_t < 67.35 **AND** PM10_t >= 49.29 **AND** NO2_t >= 57.595 **THEN** Predicted: PM10_next_1 = 43.53
3	**IF** PM10_t <108.87 **AND** PM10_t >= 84.66 **OR** MISSING **AND** CO_t < 0.815 **OR** MISSING **THEN** Predicted: PM10_next_1 = 97.41
4	**IF** PM10_t_3 >= 50.15 **AND** PM10_t < 37.335 **AND** PM10_t >= 31.995 **THEN** Predicted: PM10_next_1 = 26.4975
5	**IF** PM10_t_1 < 34.275 **AND** PM10_t < 41.165 **AND** PM10_t >= 37.335 **AND** NO2_t >= 23.35 **THEN** Predicted: PM10_next_1 = 28.37
6	**IF** PM10_t_2 < 97.21 **OR** MISSING **AND** PM10_t < 108.87 **AND** PM10_t >= 96.055 **OR** MISSING **AND** CO_t >= 0.815 **THEN** Predicted: PM10_next_1 = 100.183

The statistical parameters used in this analysis (RMSE and MAPE) present satisfactory values for the test data set: RMSE is 6.41 µg/m^3 and MAPE is 7.36 %. The running time is 19.66 seconds.

4.3 ANFIS Method

The structure of the ANFIS model is presented in Figure 2.

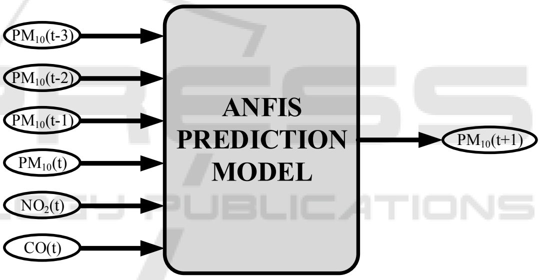

Figure 2: Structure of the ANFIS model.

The model has six inputs, namely: four PM_{10} concentrations, from current hour ($PM_{10}(t)$) to three hours ago, current value of NO_2 concentration and current value of CO concentration. The output of the model is the prediction of the next hour PM_{10} concentration – PM_{10} (t+1). The number of inputs and their granularity determine a rule base of 729 rules. The specifics of used data set imposed grid partition as FIS generating method.

This study is performed in MATLAB® where four ANFIS architectures will be used. Architecture 1 consists of Gaussian membership functions for inputs and backpropagation optimization algorithm for the training of the neural network, while architecture 2 uses the same type of membership functions but the optimization algorithm is hybrid. Similarly, architectures 3 and 4 use triangular membership functions for inputs and as optimization algorithms backpropagation and hybrid respectively.

The simulation results for each architecture, in terms of statistical parameters (RMSE, IA and MAPE), are presented in Table 6.

Table 6: Statistical parameters for ANFIS architectures.

ANFIS	RMSE[μg/m^3]	IA	MAPE[%]
Arch. 1	5.4828	0.9348	11.9772
Arch. 2	7.0774	0.9148	9.8816
Arch. 3	*5.0552*	*0.9530*	*9.2518*
Arch. 4	5.1240	0.9508	9.2606

From Table 6 it can be observed that the best values of the statistical parameters (smallest RMSE and MAPE and IA closest to 1) are obtained for the third ANFIS architecture with triangular membership functions for inputs and backpropagation as optimization algorithm. The testing error for this case is illustrated in Figure 3 and a partial view of the comparison between testing data and predicted data is presented in Figure 4.

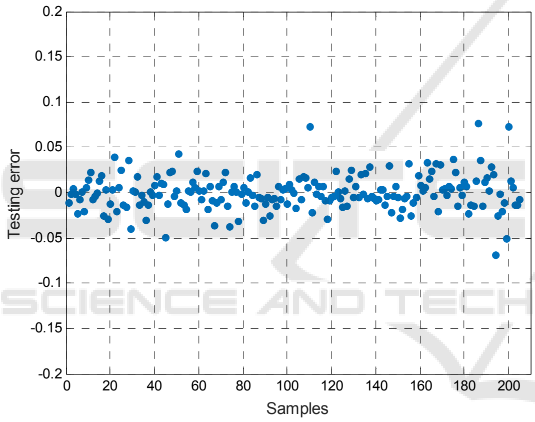

Figure 3: Testing error.

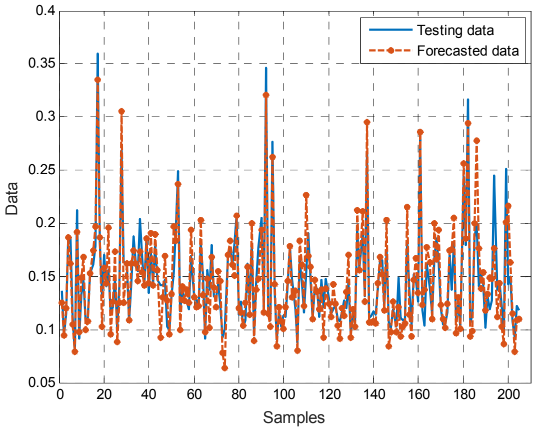

Figure 4: Comparison between testing and predicted data.

The training time for this model is around 24 hours due to the large number of rules from the fuzzy inference system.

Table 7: Comparative results.

Method	RMSE	MAPE	Running time
Data Mining	6.41	7.36	~ 1 min
ANFIS	5.06	9.25	~ 24 hours

As it can be seen in Table 7 the ANFIS method has a small value of RMSE. However, the data mining (decision tree) method has the advantage of a very small running time (which includes training, validation and testing) and a smaller value for MAPE.

5 CONCLUSIONS

Our research work investigated the potential use of two AI based prediction models: a data mining technique, decision trees, and the ANFIS model. Both models build a rule base that can be used in a knowledge base system. Decision trees generate a rule base with the extracted knowledge, while ANFIS has an internal fuzzy rule base on which the prediction is performed.

There are few research studies that compare data mining and ANFIS methods used in forecasting and our work tries to supplement this research field. Our comparative study revealed that ANFIS performs better than data mining (decision tree) method in terms of RMSE and IA, while data mining presents very small running time and smaller MAPE. The main drawback of the ANFIS method is the very long time associated to the training phase (determined by the large number of fuzzy rules, because it starts with an empty fuzzy rule base). Due to the data specifics from this case study, clustering method for generating FIS structure could not be used to diminish the number of fuzzy rules.

As future work we shall combine the two methods by using the rule base generated with the data mining method (decision tree) as the initial fuzzy rule base of the FIS structure, thus decreasing the ANFIS computational effort.

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

Ausati, S., Amanollahi, J., 2016. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$. *Atmospheric Environment*, 142, 465-474.

Bae, J. K., Kim, J., 2011. Combining models from neural networks and inductive learning algorithms, *Expert Systems with Applications*, 38, 4839-4850.

Breiman, L., Friedman, J. H., Olsen, R. A., Stone C. J., 1984. *Classification and Regression Trees*, Wadsworth. Belmont, California.

Carbajal-Hernández, J. J., Sánchez-Fernández, L. P., Carrasco-Ochoa, J. A., Martínez-Trinidad, J.F., 2012. Assessment and prediction of air quality using fuzzy logic and autoregressive models. *Atmospheric Environment*, 60, 37-50.

Domańska, D., Wojtylak, M., 2012. Application of fuzzy time series models for forecasting pollution concentrations. *Expert Systems with Applications*, 39, 7673-7679.

Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A., Samarasinghe, S, 2014. Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using neural network modelling and k-means clustering. *Atmospheric Environment*, 94, 106-116.

Hajek, P., Olej, V., 2015. Predicting common air quality index – The case of Czech Microregions. *Aerosol and Air Quality Research*, 15, 544-555.

Loya, N., Pineda, I.O., Pinto, D., Gómez-Adorno, H., Alemán, Y., 2013. Forecast of Air Quality Based on Ozone by Decision Trees and Neural Networks. *Advances in Artificial Intelligence*, LNCS vol. 7629, Springer, 97-106.

Mekanik, F., Imteaz, M. A., Talei, A., 2015. Seasonal rainfall forecasting by adaptive network-based fuzzy inference system (ANFIS) using large scale climate signals. *Climate Dynamics*, 46 (9), 3097-3111.

Mihalache, S. F., Popescu, M., Oprea, M., 2015. Particulate matter prediction using ANFIS modelling techniques. *Proceedings of the 19th International Conference on System Theory, Control and Computing, ICSTCC 2015*, 895-900.

Mishra, D., Goyal, P., Upadhyay, A., 2015. Artificial intelligence based approach to forecast $PM_{2.5}$ during haze episodes: A case study of Delhi, India. *Atmospheric Environment*, 102, 239-248.

Ong, B. T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$. *Neural Computing & Applications*, 27, 1553-1566.

Oprea, M., Mihalache, S. F., Popescu, M., 2016a. A comparative study of computational intelligence techniques applied to $PM_{2.5}$ air pollution forecasting. *Proceedings of the 6th International Conference on Computers Communications and Control, ICCCC 2016*, 103-108.

Oprea, M., Dragomir, E. G., Popescu, M., Mihalache, S. M., 2016b. Particulate Matter Air Pollutants Forecasting Using Inductive Learning Approach. *REV. CHIM.* (*Bucharest*), 67 (10), 2075-2081.

Osrodka, L., Wojtylak, M., Krajny, E., Dunal, R., Klejnowski, 2005. Application data mining for forecasting of high-level air pollution in urban-industrial area in southern Poland. *Proceedings of the 10th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*, 664-668.

Palit, A. K., Popovic, D., 2005. *Computational intelligence in time series forecasting. Theory and Engineering Applications*, Springer-Verlag. London

Perez, P., Salini, G., 2008. $PM_{2.5}$ forecasting in a large city: Comparison of three methods. *Atmospheric Environment*, 42, 8219-8224.

Polat, K., Durduran, S.S., 2012. Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM_{10}) in the city of Konya. *Neural Computing & Applications*, 21, 2153-2162.

Prasad, K., Gorai, A.K., Goyal, P., 2016. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric Environment*, 128, 246-262.

Riga, M., Tzima, F. A., Karatzas, K., Mitkas, P. A., 2009. Development and Evaluation of Data mining Models for Air Quality Prediction in Athens, Greece. In I. N. Athanasidis et al., *Information Techologies in Environmental Engineering, Environmental Science and Engineering*, Springer-Verlag, 331-344.

SAS Enterprise Miner 13.2, Reference Help, 2016.

Savić, M., Mihajlović, I., Arsić, M., Živković, Ž., 2014. Adaptive-network-based fuzzy inference system (ANFIS) model-based prediction of the surface ozone concentration. *Journal of the Serbian Chemical Society*, 79 (10), 1323-1334.

Schubert, S., Lee, T., 2011. *Time series data mining with SAS Enterprise Miner*™. SAS Global Forum 2011.

Shad, R., Mesgari, M. S., Abkar, A., Shad, A., 2009. Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *Computers, Environment and Urban Systems*, 33, 472-481.

Shahraiyni, H. T., Sodoudi, S., Kerschbaumer, A., Cubasch, U., 2015. A new structure identification scheme for ANFIS and its application for the simulation of virtual air pollution monitoring stations in urban areas. *Engineering Applications of Artificial Intelligence*, 41, 175-182.

Siwek, K., Osowski, S., 2016. Data mining methods for prediction of air pollution. *International Journal of Applied Mathematics and Computer Science*, 26 (2), 467-478.

Siwek, K., Osowski, S., Sowiński, M., 2011. Evolving the ensemble of predictors model for forecasting the daily average PM_{10}. *International Journal of Environment and Pollution*, 46 (3/4), 199-215.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60, 632-655.