Particulate Matter Air Pollutants Forecasting using Inductive Learning Approach

MIHAELA OPREA*, ELIA GEORGIANA DRAGOMIR, MARIAN POPESCU, SANDA FLORENTINA MIHALACHE
Petroleum Gas University of Ploiesti, 39 Bucuresti Blvd., Ploiesti, 100680, Romania

Particulate matter (PM) represents an important air pollutant with potential significant negative effects on human health when the level of concentration exceeds certain limits, imposed by national and international air quality standards. More accurate forecasting of such levels of PM concentrations became an important challenge nowadays, for the environmental protection specialists. Depending on the pollution sources, PM air pollutants have a variety of chemical and physical compositions. Moreover, the meteorological and geographical conditions as well as the existence of other air pollutants in the same region (to which PM can interact or which generate PM via chemical reactions), will influence the real levels of PM concentrations. Due to the high complexity of physical or chemical models that provide a more complete characterization of the PM related air pollution trends, other approximate models could be adopted. An example is an artificial intelligence based forecasting model that incorporates knowledge from the environmental expertise domain, which will guide the forecasting process. The research study presented in this paper proposes such a model, based on a machine learning approach, inductive learning, that extracts rules for guiding forecasting of the PM air pollutants concentrations levels, with better accuracy. A comparative analysis between two forecasting models based on the inductive learning algorithms, REPTree and M5P, was carried out for the forecasting of the next day PM_{10} (i.e. PM with the diameter less than 10mm) concentration level by using the last 8 days measured values. The experiments that were performed revealed that the M5P inductive learning algorithm improved the accuracy of the short-term PM_{10} concentrations levels forecasting.

Keywords: environmental protection, particulate matter air pollutant forecasting, artificial intelligence, inductive learning.

Air pollution became an important problem that need to be tackled mainly by environmental protection decision makers in order to increase the quality of life. Apart from the national and international air quality standards that impose limits for the air pollutants concentrations levels, almost all countries developed national air quality monitoring networks that provide online information regarding the air quality index over their territory and possible health effects on sensitive population (e.g. children, elderly people), when the limits are exceeded. Under this context, the forecasting of air pollution episodes would represent an efficient way to improve the quality of life in areas with higher air pollution.

Particulate matter (PM) is one of the air pollutants usually associated with poor air quality. PMs represent solid or liquid particles found in the air [1]. Their chemical and physical compositions vary widely. The main chemical characteristics of PMs are: organic or inorganic composition, and biological characteristics (e.g. bacteria, pollens, spores). The physical characteristics include particle size, optical qualities, mode of information, settling properties. Related to their size, particles can be classified into three basic categories: ultrafine particles (particles with a diameter smaller than 0.1μm), fine particles (with a diameter between 0.1 and 2.5μm), and coarse particles (with a diameter larger than 2.5μm). PM_{10} is defined as particulate matter with an aerodynamic diameter less than 10μm.

At present, the air quality standards impose limits to daily and annual average concentration levels for PM_{10} and $PM_{2.5}$. In Romania, according to [2] the upper limits are imposed for PM_{10} and they are: 50 μg/m³ - for the daily average concentration level, and 40 μg/m³ - for the annual average concentration level.

The PM origins vary from mechanical sources (sea salt, dust) to combustion (motor vehicles, industrial, fires etc.) as presented in figure 1. Particulate matter can be directly emitted or can be formed in the atmosphere when gaseous pollutants such as SO_2 and NO_x react to form fine particles [1]. The PM concentration can be also influenced by geographic location, season, day, time of day and some meteorological processes: wind, temperature, solar radiation, relative humidity, vertical mixing, precipitations, clouds, fog.

PMs are formed through different processes such as: *coagulation* – smaller particles collide and stick together, *condensation* – liquid droplets are formed when gases condense onto small solid particles, *chemical reaction* – gases react to form particles, and *cloud/fog processes* – gases dissolve in a water droplet and chemically react and when the water evaporates particles remain.

The composition of particulate matter is given by its sources and may be changed by different chemical reactions (see e.g. the research studies described in [3, 4]). For example, in coastal areas, PM may present high sulfate and low chloride contents. The sulfate comes from atmospheric oxidation of sulfur dioxide to form nonvolatile ionic sulfate, whereas some chloride may be lost from the solid aerosol as volatile HCl [5]:

$$2SO_2 + O_2 + 2H_2O \rightarrow 2H_2SO_4$$

$$H_2SO_4 + 2NaCl\,(particulate) \rightarrow$$

$$\rightarrow Na_2SO_4\,(particulate) + 2HCl$$

In the presence of basic air pollutants, such as ammonia or calcium oxide, the sulfuric acid reacts to form salts [6]:

* email: mihaela@upg-ploiesti.ro

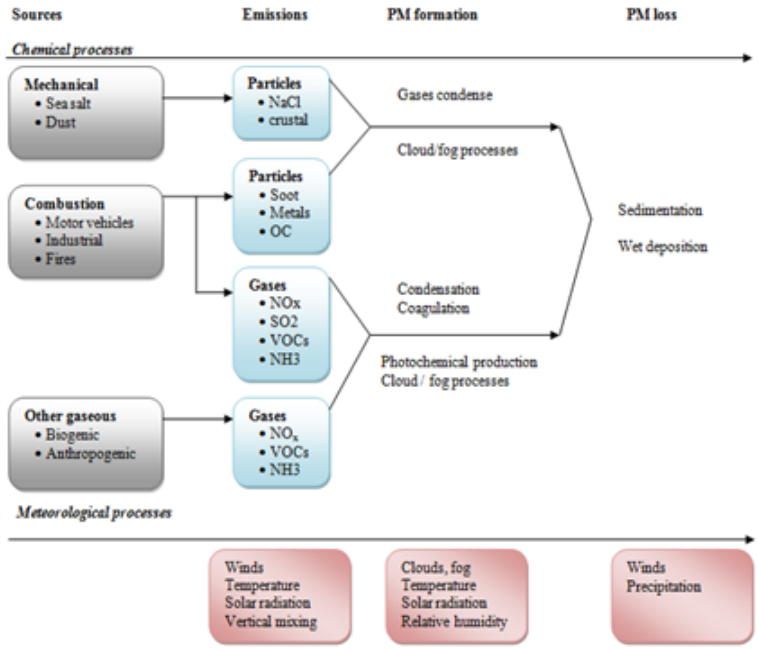

Fig. 1. Particulate matter chemistry (adapted after [1])

$$H_2SO_4(droplet) + 2NH_3(g) \rightarrow (NH_4)_2SO_4(droplet)$$
$$H_2SO_4(droplet) + CaO(s) \rightarrow CaSO_4(droplet) + H_2O$$

Water is lost from these droplets when low-humidity is present and a solid aerosol is formed.

Another common component of particulate matter is the nitrogen in ammonium and nitrate salts. Particulate ammonium nitrate and ammonium chloride are produced by the following reversible reactions [6]:

$$NH_3(g) + HNO_3(g) \rightleftarrows NH_4NO_3(s, particulate)$$
$$NH_3(g) + HCl(g) \rightleftarrows NH_4Cl(s, particulate)$$

These types of reactions constitute an important process associated to particulate matter formation.

The population exposure to high levels of PM concentration may lead to important health problems. This is why it is necessary to develop more accurate forecasting models for PM concentration level and to integrate them as software tools on web-applications (e.g. under the national air quality monitoring network) which will inform, in real time, the public and the governmental authorities, when PM air pollution episodes arise.

Various forecasting PM models were experimented so far, starting from the physical and chemical models (complex models, but most complete ones, requiring complete real data, which are not always available), statistical models (e.g. ARMA, ARIMA, LR, MLR), and artificial intelligence based models (e.g. computational intelligence models as artificial neural networks and neuro-fuzzy models [7-12], data mining models [13-17]). The inductive learning model (i.e. a machine learning model) is an artificial intelligence based model that provides predictive rules guiding the PM concentration level forecasting result. The approach used in this research study combines the Principal Component Analysis (PCA), a statistical technique that is applied to the detection of the environmental parameters (air pollutants and meteorological parameters) that have the most important influence on the PM_{10} concentration level, with inductive learning, in order to increase the forecasting accuracy.

Experimental part

The Inductive Learning Approach

Inductive learning is an automatic knowledge extraction technique that generates a structured knowledge representation [18], based on a set of input examples formed by the values of some attributes and the class it belongs. It can be used to automatically discover knowledge, usually, under the form of IF-THEN rules, from historical databases, and it can be applied to any expertise domain, in particular, to discover predictive rules for PM air pollutant concentration levels forecasting by using historical databases with various parameters: meteorological (e.g. wind speed, air temperature, relative humidity, solar radiation), the past concentration levels of PM and other PM related air pollutants (e.g. SO_2, NO_x).

Currently, there are two types of approaches for knowledge generation [19]: (1) approaches that will generate a decision tree and will perform rule induction from the decision tree, and (2) approaches that will generate directly the rules from the examples set.

The first type of approach includes methods that use a dividing and conquering algorithm to divide the original examples set into some subsets, based on specific strategies (e. g. information gain, entropy), that will generate a decision tree which represents the knowledge generalized from the given examples set. Thus, it can be used to analyze situations that are not explicitly described in the input set.

Under the second type of approach (named the covering approach), the inductive learning method identifies groups of attributes uniquely shared by the examples set. Furthermore, it forms IF – THEN rules. This is also a useful approach to discover interesting data relations and to acquire structured knowledge.

The knowledge derived through inductive learning is included in a knowledge base that is used later to solve different problems (e.g. PM air pollutant concentration level forecasting).

Various inductive learning algorithms were developed [17, 19-21] such as: CART, ID3 and different version of it (C4.5, C5.0), RuleQuest, M5, which build a decision tree in order to generate domain knowledge.

In our research study we have used two inductive learning algorithms, REPTree and M5P.

The REPTree algorithm builds a regression tree based on the principle of calculating the information gain with entropy and reducing the error arising from variance [22]. It only sorts values for numeric attributes once. This algorithm uses the method from C4.5 and the basic REP (Reduce Error Prunning).

The M5P algorithm is an improvement of the Quinlan's M5 algorithm for inductive trees models. It is a combination between conventional decision trees and the linear regression. The linear functions are used at the tree's nodes. Instead of maximizing the information gain, at each node is used a splitting criterion that minimizes the intra-subset variation in the class values down each branch [22]. The second step, after building the tree, is the pruning process, followed by the smoothing procedure applied to avoid sharp discontinuities. The M5P model tree inducer generates accurate classifiers, particularly when most of the attributes are numeric.

Air Pollution Forecasting

Air pollution forecasting involves the concentrations levels prediction of the most important air pollutants specific to a certain region: carbon monoxide, sulfur dioxide, nitrogen monoxide, nitrogen dioxide, nitrogen oxides, particulate matter, ozone, benzene, lead, polycyclic aromatic hydrocarbons. These values are directly influenced by meteorological factors (such as wind speed, wind direction, air temperature, relative humidity etc), seasons, geographical area (topography, buildings etc) or industrial activity in the analyzed region.

The forecast can be done for short term (1 hour, 1 day, 1 month) or long term (1 or more years), depending mainly on the size of the database that stores the recorded values of the past measurements.

Air pollution forecasting for a certain air pollutant concentration can be guided using heuristic rules, which take into account physical phenomena and chemical reactions that occur in the troposphere and incorporate the domain human expert knowledge. Examples of such heuristic rules are given in table 1. The rules can be applied in the first step of the air quality knowledge generation methodology (proposed in [23]), significant input parameters selection, in order to verify the results obtained for the selection of the parameters that can influence the concentration level of a particular air pollutant.

In our research study we have used the following methodology [23] for the generation of the PM air pollutant knowledge base.

The air quality knowledge generation methodology

The input of the algorithm is represented by the database with air quality parameters measurements, and the output is a knowledge base.

The steps of the knowledge generation methodology are:

- selection of significant input parameters for predicting a certain variable using a proper selection method (e.g. Principal Component Analysis - PCA) and create another database only with these parameters;
- data preprocessing;
- applying a knowledge generation algorithm;
- building a new knowledge base with the new generated knowledge or appending it to the existing one.

The quality of the extracted environmental knowledge depends on the data sets resulted after the data pre-processing step [24].

Forecasting Method and Data Set

The proposed PM forecasting method has the following main steps:

- select the type of PM air pollutant that will be forecasted (i.e. PM_{10}, $PM_{2.5}$, $PM_{0.1}$) and the forecasting horizon (e.g. next hour, next 24 h);
- choose the databases (DB) with time series of meteorological and air pollutants concentrations measurements (e.g. hourly or daily average) and create a new PM specific database (PM-DB);
- generate from PM-DB the PM knowledge base (PM-KB) with predictive rules (i.e. apply the air quality knowledge generation methodology – inductive learning approach);
- perform PM air pollutant concentration level forecasting by using the heuristic rules from PM-KB.

The two inductive learning algorithms, REPTree and M5P, are applied in step 3 of the PM predictive knowledge generation methodology. These algorithms are implemented in Weka (Waikato Environment for Knowledge Analysis), a free data mining software tool [25], based on java, which was used in the experiments that were carried out.

In our case study, the data set included daily time series data for the following air pollutants: sulfur dioxide ($\mu g/m^3$), nitrogen monoxide ($\mu g/m^3$), carbon monoxide (mg/m^3), nitrogen oxides, nitrogen dioxide ($\mu g/m^3$), ($\mu g/m^3$), particulate matter ($\mu g/m^3$), ozone ($\mu g/m^3$), o-xylene ($\mu g/m^3$), m-xylene ($\mu g/m^3$), benzene ($\mu g/m^3$), toluene ($\mu g/m^3$), p-xylene ($\mu g/m^3$), 1,3 butadiene ($\mu g/m^3$), ethyl-benzene ($\mu g/m^3$), as well as for the meteorological parameters: temperature (°C), relative humidity (%), solar radiation (W/m^2), atmospheric pressure (mbar), wind direction (deg), wind speed (m/s), precipitations (mm), over a period of 27 months (from January 2009 to December 2009, from January 2011 to December 2011, and from January 2015 to April 2015). The data were recorded at the Ploie°ti monitoring stations integrated in the Romanian Air Quality Monitoring Network [2] and they were preprocessed in order to be used by the Weka inductive learning algorithms.

In the first step of the air quality knowledge generation methodology, for these time series data it was applied a Principal Component Analysis. The purpose of PCA was to identify the subset of attributes that are most relevant in solving a particular problem. This allows reducing the size

Rule No	Heuristic Rules
1	**IF** forecast parameter is ozone **THEN** analyzed parameters can be ozone, solar radiation, wind speed, Volatile Organic Compounds, nitrogen oxides etc.
2	**IF** forecast parameter is particulate matter **THEN** analyzed parameters can be particulate matter, wind direction, wind velocity, sulfur dioxide, temperature, precipitations etc.
3	**IF** forecast parameter is nitrogen oxides **THEN** analyzed parameters can be nitrogen monoxide, nitrogen dioxide, nitrogen oxides, temperature, solar radiation, Volatile Organic Compounds etc.
4	**IF** forecast parameter is sulfur dioxide **THEN** analyzed parameters can be sulfur dioxide, aromatic hydrocarbons, wind velocity etc.

Table 1
EXAMPLE OF HEURISTIC RULES

Principal Component	Eigenvalue	Variance	Cumulative
PC1	6.11823	0.2781	0.2781
PC2	5.44853	0.24766	0.52576
PC3	2.70879	0.12313	0.64889
PC4	2.09474	0.09522	0.7441
PC5	1.57799	0.07173	0.81583

Table 2
THE PCA EXPERIMENTAL RESULTS

The attribute	The attribute description
PM_{10_i} SO_{2_i} NO_{2_i} $Temp\_i$ $Humid\_i$	The values of the specified parameters concentration, recorded *i* days before, i=1,8
$PM_{10\_next\_j}$	The PM_{10} concentration value forecast for *j* days ahead

Table 3
THE PM_{10} DATABASE DESCRIPTION

of the input vector and inductive learning algorithms can run more efficiently and with a reduced execution time.

The experiments were carried out for PM_{10} air pollutant concentration level forecasting. The selection of the most significant air parameters involved in PM_{10} forecasting was performed via the PCA technique. The input vector was first normalized and afterwards, the components that have the least contribution to the variation in the data set were eliminated. As a result, 5 principal components were used as inputs in the PM_{10} knowledge generation models. The corresponding eigenvalue, variance and cumulative values were determined for each principal component (PC), and are shown in table 2. The principal component analysis extracted 5 PC with initial eigenvalues greater than unity. Approximately 81% of the variance between the data points can be explained by these 5 PC.

PCA selected sulfur dioxide (SO_2), nitrogen dioxide (NO_2), particular matter (PM_{10}), temperature and relative humidity as the most important parameters for the PM_{10} forecast. Wind speed was not taken into consideration, mainly because in this period all the recorded values were below 0.1 m/s. The air pollutants that were selected by PCA are specific to Ploie°ti, which is an industrial city with the petrochemical and chemical industry as main contributor to environmental pollution.

The PM_{10} time series forecasting problem formulation

According to the PCA experiments, the PM_{10} concentration level at time *i+1* depends on the *k* past values for PM_{10}, SO_2, NO_2 air pollutants concentrations and the following meteorological parameters: temperature and humidity. The selection of these air parameters is validated by the physical and chemical characteristics of PM_{10} along with domain experts' heuristics rules: mainly low wind speed, rare days with precipitation and stationary temperature inversions are responsible for high PM_{10} loads in the Ploie°ti city.

The formal representation of the PM_{10} time series forecasting model is given by equation (1), in a similar way to the forecasting model described in [23, 26].

$$x_{i+p} = f(x_i, x_{i-1}, \ldots, x_{i-k}, y_i, y_{i-1}, \ldots, y_{i-k}, z_i, z_{i-1}, \ldots, z_{i-k}, w_i, w_{i-1}, \ldots, w_{i-k}, q_i, q_{i-1}, \ldots, q_{i-k}) \quad (1)$$

where:

-*k* represents the number of the past values taken into consideration for prediction;

-*p* represents the time window used in the forecast (the forecasting horizon, e.g. the number of hours or days ahead);

-x_{i+1} is the daily PM_{10} level at time *i+1*;

-x_i, x_{i-1},..., x_{i-k} are the daily PM_{10} level recorded at the time *i, i-1,..., i-k*;

-y_i, y_{i-1},..., y_{i-k} are the daily temperature value recorded at the time *i, i-1,..., i-k*;

-z_i, zi_{-1},..., z_{i-k} are the daily humidity level recorded at the time *i, i-1,..., i-k*;

-w_i, w_{i-1},..., w_{i-k} are the daily SO_2 level recorded at the time *i, i-1,..., i-k*;

-q_i, q_{i-1},..., q_{i-k} are the daily NO_2 level recorded at the time *i, i-1,..., i-k*.

In our experiments, we have considered *k* to be 8 (in order to capture as much as possible the physical and chemical PM_{10} formation processes), and p=1, 2 and 3 as the next *p* days PM_{10} value can be determined by knowing the previous 8 days recorded values for PM_{10}, SO_2, NO_2, air temperature and relative humidity. The attributes of the new database built with these atmospheric parameters are presented in table 3.

The representation of PM_{10} concentrations levels evolution and of the PCA selected atmospheric parameters is provided for each annual time series, in order to extract or validate the correlations between the analyzed parameters.

The evolution of the annual PM_{10} concentration (during 2011) compared with the concentrations of SO_2 and NO_2 air p[illegible] h the

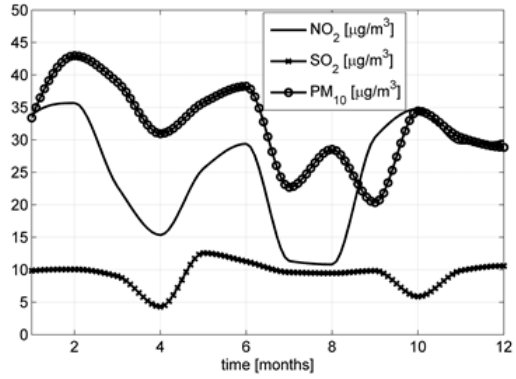

Fig. 2. Annual concentration evolution of PM_{10} against SO_2 and NO_2

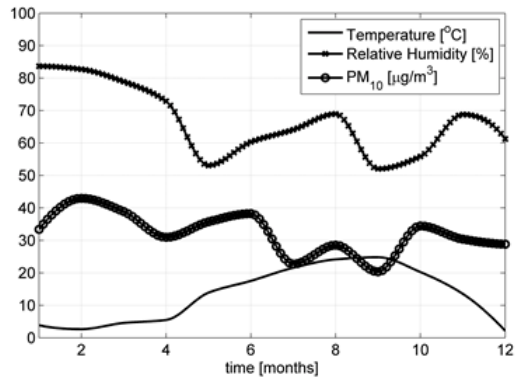

Fig. 3. Annual concentration evolution of PM_{10} against air temperature and relative humidity

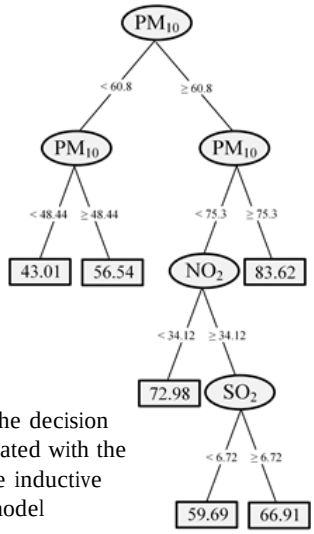

Fig. 4. The decision tree generated with the REPTree inductive model

air temperature and relative humidity is depicted in figure 3.

The experiments were realized under the Weka data mining toolkit, by using the REPTree and M5P inductive algorithms, for different values of *p* (from 1 to 3 days ahead forecast). The most representative decision trees (for 1 day before) are presented in the next section.

Results and discussions

The application of the proposed forecasting method to PM_{10} concentration level forecasting provides for each inductive learning algorithm (i.e. inductive forecasting model) that was run under Weka software, a decision tree and a set of heuristic predictive rules. These are discussed in the following sections.

The REPTree inductive learning algorithm run

The decision tree generated by the REPTree algorithm run for PM_{10} concentration forecast is shown in figure 4. In the root node, it is selected as the most relevant parameter, the PM_{10} concentration recorded 1 day before. After this selection, the algorithm chooses the NO_2 concentration value measured 8 days before, and the 7 days before concentration for SO_2.

Table 4
EXAMPLES OF HEURISTIC RULES FOR THE REPTREE MODEL

Rule No	Heuristic predictive rules
1	**IF** PM_{10_1} <48.44 **THEN** $PM_{10\_next\_1}$=43.01
2	**IF** PM_{10_1} > 48.44 **AND** PM_{10_1} <60.8 **THEN** $PM_{10\_next\_1}$=56.54.
3	**IF** PM_{10_1} >=60.8 **AND** PM_{10_1} <75.3 **AND** NO_{2_8} < 34.12 **THEN** $PM_{10\_next\_1}$=72.98.
4	**IF** PM_{10_1} >60.8 **AND** PM_{10_1} <75.3 **AND** NO_{2_8} > =34.12 **AND** SO_{2_7}<6.72 **THEN** $PM_{10\_next\_1}$=59.69.
5	**IF** PM_{10_1} >60.8 **AND** PM_{10_1} <75.3 **AND** NO_{2_8} > =34.12 **AND** SO_{2_7}>=6.72 **THEN** $PM_{10\_next\_1}$=66.91.
6	**IF** PM_{10_1} >= 75.3 **THEN** $PM_{10\_next\_1}$ = 83.62

Table 4 presents some examples of knowledge extracted from the generated decision tree by the REPTree inductive model. The knowledge has the form of IF-THEN rules. The model took into consideration for the next day PM_{10} concentration level forecasting, the PM_{10} concentration recorded 1 day before, the NO_2 concentration for 8 days before and the 7 days before for the SO_2 concentration.

The M5P inductive learning algorithm run

The decision tree generated by the M5P algorithm run for PM_{10} concentration forecast is shown in figure 5. PM_{10_1}, PM_{10_6}, SO_{2_7} and SO_{2_3} are considered by the M5P inductive model as the most important parameters for the PM_{10} concentration level forecast.

Eight linear models (LM) were built during the M5P decision tree generation, which are applied in different cases to determine the PM_{10} next day concentration level. The linear regression functions were generated based on the atmospheric parameters values recorded *k* days before and they are shown in the relations (2)-(9). For example, the LM1 linear regression function calculates the 1 h ahead PM_{10} concentration level based on the PM_{10} values recorded in the 5 previously days as well as the 8 days ahead recorded concentrations for SO_2, NO_2, and temperature. Along these, the linear model took into consideration 7, 6, 3 and 2 days ahead SO_2 measured concentrations. This selection can be explained based on the PM_{10} chemical processes: PM_{10} can be formed in the atmosphere when gaseous pollutants

$$\mathbf{LM1:}\quad PM_{10\_next\_1} = 0.171\cdot SO_{2\_8} - 0.0942\cdot SO_{2\_7} + 0.2583\cdot SO_{2\_6} + 0.1806\cdot SO_{2\_3} - 0.0518\cdot SO_{2\_2} - 0.0593\cdot NO_{2\_8} + 0.0495\cdot PM_{10\_5} - 0.0882\cdot PM_{10\_4} + 0.1338\cdot PM_{10\_3} + 0.4399\cdot PM_{10\_1} - 0.2134\cdot Temp\_8 + 20.4968 \quad (2)$$

$$\mathbf{LM2:}\quad PM_{10\_next\_1} = 0.2185\cdot SO_{2\_8} - 0.0281\cdot SO_{2\_7} + 0.2583\cdot SO_{2\_6} + 0.1806\cdot SO_{2\_3} - 0.0518\cdot SO_{2\_2} - 0.0758\cdot NO_{2\_8} + 0.0495\cdot PM_{10\_5} - 0.0882\cdot PM_{10\_4} + 0.1338\cdot PM_{10\_3} + 0.4399\cdot PM_{10\_1} - 0.2134\cdot Temp\_8 + 20.8078 \quad (3)$$

$$\mathbf{LM3:}\quad PM_{10\_next\_1} = -0.2468\cdot SO_{2\_7} + 0.2583\cdot SO_{2\_6} + 0.1806\cdot SO_{2\_3} - 0.0892\cdot SO_{2\_2} + 0.0495\cdot PM_{10\_5} - 0.0882\cdot PM_{10\_4} + 0.0994\cdot PM_{10\_3} + 0.3954\cdot PM_{10\_1} - 0.8677\cdot Temp\_8 + 31.5464 \quad (4)$$

$$\mathbf{LM4:}\quad PM_{10\_next\_1} = -0.1407\cdot SO_{2\_8} + 1.485\cdot SO_{2\_7} + 0.1398\cdot SO_{2\_6} + 0.4668\cdot SO_{2\_3} - 0.1068\cdot SO_{2\_2} - 0.0728\cdot PM_{10\_7} + 0.0268\cdot PM_{10\_5} - 0.232\cdot PM_{10\_4} + 0.341\cdot PM_{10\_1} - 0.1155\cdot Temp\_8 + 43.5034 \quad (5)$$

$$\textbf{LM5:}\quad \begin{aligned} PM_{10\_next\_1} &= -0.2422\cdot SO_{2\_8} + 1.6027\cdot SO_{2\_7} + 0.1398\cdot SO_{2\_6} + 0.4141\cdot SO_{2\_3} - \\ &- 0.1068\cdot SO_{2\_2} - 0.0728\cdot PM_{10\_7} + 0.0268\cdot PM_{10\_5} - 0.232\cdot PM_{10\_4} + \\ &+ 0.341\cdot PM_{10\_1} - 0.1155\cdot Temp\_8 + 44.7757 \end{aligned} \quad (6)$$

$$\textbf{LM6:}\quad \begin{aligned} PM_{10\_next\_1} &= -0.1407\cdot SO_{2\_8} + 1.2656\cdot SO_{2\_7} + 0.1398\cdot SO_{2\_6} + 0.0977\cdot SO_{2\_3} - \\ &- 0.1068\cdot SO_{2\_2} - 0.0728\cdot PM_{10\_7} + 0.0268\cdot PM_{10\_5} - 0.2759\cdot PM_{10\_4} + \\ &+ 0.341\cdot PM_{10\_1} - 0.1155\cdot Temp\_8 + 51.4832 \end{aligned} \quad (7)$$

$$\textbf{LM7:}\quad \begin{aligned} PM_{10\_next\_1} &= -0.1407\cdot SO_{2\_8} + 0.6163\cdot SO_{2\_7} + 0.1398\cdot SO_{2\_6} + 0.0977\cdot SO_{2\_3} - \\ &- 0.1068\cdot SO_{2\_2} - 0.0738\cdot PM_{10\_7} + 0.0654\cdot PM_{10\_5} - 0.1577\cdot PM_{10\_4} + \\ &+ 0.3436\cdot PM_{10\_1} - 0.1155\cdot Temp\_8 + 58.1985 \end{aligned} \quad (8)$$

$$\textbf{LM8:}\quad \begin{aligned} PM_{10\_next\_1} &= -0.2731\cdot SO_{2\_8} + 0.4884\cdot SO_{2\_7} + 0.1398\cdot SO_{2\_6} + 0.0977\cdot SO_{2\_3} - \\ &- 0.1068\cdot SO_{2\_2} + 0.7752\cdot SO_{2\_1} - 0.0973\cdot NO_{2\_3} - 0.057\cdot PM_{10\_7} + \\ &+ 0.0268\cdot PM_{10\_5} - 0.723\cdot PM_{10\_4} + 0.3307\cdot PM_{10\_2} + 0.3259\cdot PM_{10\_1} - \\ &- 0.1155\cdot Temp\_8 + 71.1937 \end{aligned} \quad (9)$$

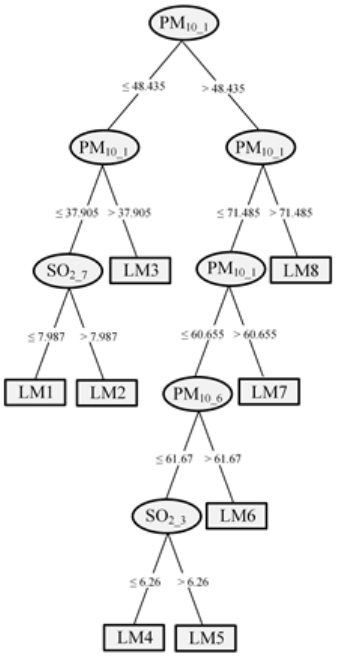

Fig. 5. The decision tree generated with the M5P inductive model

such as SO_2 and NO_x react to form fine particles influenced by temperature.

Examples of knowledge extracted by the M5P inductive model are shown in table 5. The knowledge has the heuristic rules form and they guide the PM_{10} forecasting, by choosing in each case, the corresponding linear regression model.

Comparative analysis of the forecasting inductive models

A comparative analysis between the results obtained for PM_{10} concentration level forecasting with the two inductive learning models, REPTree and M5P, was performed. The forecasting performance was measured by three statistical parameters: the correlation coefficient (R), given by equation (10), mean absolute error (MAE), given by equation (11), and root mean squared error (RMSE), given by equation (12). The time taken to build the model and the number of rules / size of the decision tree were also considered for the results comparison.

$$R = \frac{\sum_{i=1}^{n} y_i \hat{y}_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} \hat{y}_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n_y\right)\left(\sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i\right)^2 / n_{\hat{y}}\right)}} \quad (10)$$

Table 5

EXAMPLES OF HEURISTIC RULES FOR THE M5P MODEL

Rule no	Heuristic rules
1	**IF** PM_{10_1} <= 48.43 **AND** PM_{10_1} <= 37.90 **AND** SO_{2_7}<= 7.98 **THEN** LM1
2	**IF** PM_{10_1} <= 48.43 **AND** PM_{10_1} <= 37.90 **AND** SO_{2_7}> 7.98 **THEN** LM2
3	**IF** PM_{10_1} <= 48.43 **AND** PM_{10_1} > 37.90 **THEN** LM3
4	**IF** PM_{10_1} > 48.43 **AND** PM_{10_1} <= 60.65 **AND** PM_{10_6}<= 61.67 **AND** SO_{2_3}<=6.26 **THEN** LM4
5	**IF** PM_{10_1} > 48.43 **AND** PM_{10_1} <= 60.65 **AND** PM_{10_6}<= 61.67 **AND** SO_{2_3}> 6.26 **THEN** LM5
6	**IF** PM_{10_1} > 48.43 **AND** PM_{10_1} <= 60.65 **AND** PM_{10_6}> 61.67 **THEN** LM6
7	**IF** PM_{10_1} > 60.65 **AND** PM_{10_1}<= 71.48 **THEN** LM7
8	**IF** PM_{10_1} > 71.48 **THEN** LM8

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \quad (11)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (12)$$

where y_i are the actual values and $\hat{y}_i$ are the forecasted values.

Table 6
EXPERIMENTAL RESULTS

Statistical parameter	1 day ahead		2 days ahead		3 days ahead	
	M5P	REPTree	M5P	REPTree	M5P	REPTree
R	***0.9301***	0.8651	0.886	0.8543	0.8774	0.6562
MAE	***4.644***	6.473	6.231	6.6019	6.5278	9.3466
RMSE	***6.3307***	8.6028	7.9799	8.9402	8.3877	12.7446
Number of Rules/ Size of the tree	8	6	3	13	14	9
Time taken to build the model	0.03 s	0.01 s	0.05s	0.02	0.12	0.07

Table 6 synthesizes the experimental results obtained for the 1 day, 2 days, and 3 days ahead PM_{10} concentration forecasting experiments.

We can notice that the correlation coefficient, the mean absolute error and root mean squared error values are inverse proportional with the value of *p*; the statistical parameters have smaller values for a time window (forecasting horizon) greater than 1. For 1 day ahead of the PM_{10} concentration forecasting, the M5P inductive model has the best correlation coefficient, 0.9301, while in the case of the REPTree model it is 0.8651; for 2 days forecasting horizon, the coefficient correlation values are 0.8860 (M5P) and 0.8543 (REPTree), respectively, and for 3 days, the correlation coefficient is 0.8774 (M5P) and 0.6562 (REPTree).

The lowest MAE was obtained by the 1 day ahead M5P model (4.644) and the greatest value for the 3 days ahead REPTree model (9.3466). The M5P inductive model has better results (6.3307) when the RMSE parameter is compared. The REPTree model is built in 0.01 s for the 1 day ahead PM_{10} air pollutant forecasting case, a better time comparing with the M5P model.

Even if the number of the fitted parameters is greater for the M5P model (i.e. 4) than for the REPTree model (i.e. 3), the M5P algorithm over performs the REPTree algorithm, from the statistical point of view.

In conclusion, for our case study, the M5P inductive model was more suitable for PM_{10} concentration level forecasting based on heuristic predictive rules, providing a better forecasting accuracy.

Conclusions

The paper presented an inductive forecasting model for PM air pollutant concentration. Being a machine learning approach, inductive learning generates decision trees and extracts heuristic predictive rules for guiding the PM concentration levels forecasting process, with a better accuracy. A comparative study between two inductive forecasting models based on the inductive learning algorithms, REPTree and M5P, was carried out for the forecasting of the next 1 day, 2 days and 3 days PM_{10} concentrations levels, in the Ploieºti city, by using the last 8 days measured values of the atmospheric parameters that were selected by the PCA technique (SO_2, NO_2, air temperature and relative humidity).

The experimental results revealed that both inductive learning models provided a good overall PM_{10} forecasting performance, thus, recommending the use of inductive learning approach in PM air pollutant concentration forecasting. The M5P inductive learning algorithm improved the accuracy of the short-term PM_{10} concentrations levels forecasting.

Acknowledgment: 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.US ENVIRONMENTAL PROTECTION AGENCY, Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program, 2003.
2.*** http://www.calitateaer.ro - the Romanian National Air Quality Monitoring Network.
3.CHEN, H.-W., CHEN, W.-Y., CHANG, C.-N., CHUANG, Y. C., Aerosol and Air Quality Research, 13, No. 2, 2013, p. 699.
4.OANH, N.T.K., THIANSATHIT, W., BOND, T., SUBRAMANIAN, R., WINIJKUL, R. E., Atmospheric Environment, 44, No. 1, 2010, p. 15.
5.MANAHAN S.E., Fundamentals of Environmental Chemistry, Third Edition, CRC Press, 2011.
6.MANAHAN S.E., Environmental Chemistry, Ninth Edition, CRC Press, 2009.
7.CHERKASSKY, V., KRASNOPOLSKY, V., SOLOMATINE, D.P., VALDES, J., Neural Networks, vol. 19, 2006, p. 113.
8.GRIVAS, G., CHALOULAKOU, A., Atmospheric Environment, 40, Is. 7, 2006, p. 1216.
9.KURT, A., OKTAY, A. B., Expert Systems with Applications, 37, Is. 12, 2010, p. 7986.
10.SHIVA NAGENDRA, S M., KHARE, M., Ecological Modelling, 190, Is. 1-2, 2006, p. 99.
11.DRAGOMIR, E G., OPREA, M., Buletinul Institutului Politehnic din Iaºi, Tome LX (LXIV), Fasc. 3/ 2014, Automatic Control and Computer Science Section, p. 265.
12.DRAGOMIR, E. G., OPREA, M., 22nd International conference NDES2014, Communications in Computer and Information Science, vol. 438, 2014, p. 387.
13.CACERES, S.Z., LOPEZ, J.T., Proceedings of EEEAD2013, 2013, p. 107.
14.DONG, M., YANG, D., KUANG, Y., HE, D., ERDAL, S., KENSKI, D., Expert Systems with Applications, 36, Is. 5 2009, p. 9046.
15.PASERO, E., MONTUORI, A., MONIACIA, W., RAIMONDO, G., Proceedings of iEMS, 2008, p. 1110.
16.SFETSOS, A., VLACHOGIANNIS, D., Proceedings of iEMS, 2008, p. 1727.
17.ZHAO, Y., HASAN, Y. A., International Journal of Advanced Computer Science and Applications, 4, No. 5, 2013, p. 21.
18.WANG, Y., WITTEN, I. H., Proc. Poster papers of ECML1997, p. 128.
19.MITCHELL, T.M., Machine Learning, McGraw-Hill, Boston, 1997.
20.QUINLAN, R. J., Proceedings AI'92 (Adams & Sterling, Eds.), 1992, p. 343.
21.QUINLAN, J.R, Machine Learning, 16, Is. 3, 1993, p. 235.
22.ANDINA, D., PHAM D. T. (eds.), Computational Intelligence for Engineering and Manufacturing, Springer, 2007.
23.PETRE (DRAGOMIR), E. G., Automated system based on intelligent agents for air quality monitoring and analysis (In Romanian), PhD Thesis, Petroleum–Gas University of Ploieºti, Romania, 2014.
24.GIBERT, K., IZQUIERDO, J., HOLMES, G., ATHANASIADIS, I., COMAS, J., SÀNCHEZ-MARRÈ, M., Proceedings of iEMS 2008, p. 1937.
25.*** http://www.cs.waikato.ac.nz/ ml/weka/
26.OPREA, M., DRAGOMIR, E.G., MIHALACHE, S.F., POPESCU, M., Methods for the assessment of air pollution with particulate matter to children's health (in Romanian) (Iordache, S., Dunea D. (eds.)), *Chap.* Prediction methods and techniques for PM2.5 concentration in urban environment (in Romanian), MatrixRom, 2014, p. 387.

Manuscript received:2.10.2015