Development and Evaluation of Data Mining Models for Air Quality Prediction in Athens, Greece

Marina Riga[1], Fani A. Tzima[1], Kostas Karatzas[2], and Pericles A. Mitkas[1]

mriga@issel.ee.auth.gr, fani@olympus.ee.auth.gr, kkara@eng.auth.gr, mitkas@auth.gr

[1]Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece

[2]Dept. of Mechanical Engineering, Aristotle University of Thessaloniki, Greece

Abstract

Air pollution is a major problem in the world today, causing undesirable effects on both the environment and human health and, at the same time, stressing the need for effective simulation and forecasting models of atmospheric quality. Targeting this adverse situation, our current work focuses on investigating the potential of data mining algorithms in air pollution modeling and short-term forecasting problems. In this direction, various data mining methods are adopted for the qualitative forecasting of concentration levels of air pollutants or the quantitative prediction of their values (through the development of different classification and regression models respectively) in five locations of the greater Athens area. An addi-

tional aim of this work is the systematic assessment of the quality of experimental results, in order to discover the best performing algorithm (or set of algorithms) that can be proved to be significantly different from its rivals. Obtained experimental results are deemed satisfactory in terms of the aforementioned goals of the investigation, as high percentages of correct classifications are achieved in the set of monitoring stations and clear conclusions are drawn, as far as the determination of significantly best performing algorithms is concerned, for the development of air quality (AQ) prediction models.

1 Introduction

Air pollution is a major problem in the world today, causing undesirable effects on both the environment and human health. The successful simulation and forecasting of atmospheric quality has, thus, become a great concern for city authorities worldwide. Furthermore, evident is the need for accurate and on-time information provision to governmental entities and citizens.

In this setting, superior organizations, such as the World Health Organization (WHO) and the European Union (EU), have introduced Air Quality (AQ) Standards, defining guidelines (limit values, corresponding margins of tolerance and information/ alert thresholds per pollutant) that each country should follow, in order to ensure the protection of human health and the ecosystem. In the same direction, the AQ community has investigated different approaches and developed efficient methodologies for dealing with the processes of modeling air quality and predicting air pollutants' concentration levels.

The efforts invested in designing and developing AQ modeling schemes have resulted in a set of requirements that every such system should meet: it should be able to (i) properly analyze and extract new information from available data and (ii) dynamically learn through previously obtained knowledge or newly entered data, in order to progressively optimize the accuracy of its predictions. However, the complexity and the heterogeneity of environmental data, the non-linearity of their correlations, in addition to the demands for computational efficiency and dynamic learning pose considerable obstacles in the effective implementation of AQ modeling schemes. Another characteristic aspect of the domain of AQ that further complicates the situation is the fact that the problem is often described by insufficient or bad quality data, with multiple missing values for the prediction variables (Karatzas et al. 2008).

The adverse situation described above, with its diverse characteristics and difficulties, forms an ideal setting for the application of Data Mining (DM) methods to "extract hidden patterns and relationships, by taking into account large amounts of data" (Goebel and Gruenwald 1999). This fact gave us the motivation to investigate their applicability comprehensively, in the real-world application domain of forecasting later ozone concentration levels in the Athens urban area. Our basic objective was the development of DM models and the evaluation of their performance using a robust methodology appropriate for the task at hand. We consider this methodology, that is based on a comprehensive study of the related literature in both the domain of AQ forecasting and DM in general, as the basic contribution of our current work.

The remainder of the paper is structured as follows: Section 2 reviews related approaches in the domain of AQ forecasting and provides some insights on the evaluation methodologies traditionally used in this line of studies. Section 3 starts by presenting the study area and the utilized air quality measurements and continues with the description of the data preprocessing methodology adopted and the overall experimental setting for the development of AQ prediction models. The section concludes with the report of the integrated statistical methodology adopted for the evaluation of results of the models' performance. Finally, Section 4 outlines the most significant results obtained from our current experimental investigation of the potential of DM methods in the problem of AQ forecasting, reinforcing them with the adopted statistical evaluation. We conclude this paper with a discussion on the future aspects of this work.

2 Related Work

According to bibliographic research, several scientific papers have been published in the domain of AQ modeling and forecasting, differing in one or more aspects of their basic approach to the problem. These differences may concern: (i) the initial training set and the corresponding variables, (ii) the algorithms adopted during the modeling phase, (iii) the desired output (classification or regression) models, and/or (iv) the evaluation process adopted for assessing experimental results.

Regarding the second aspect, a careful examination of the literature allows one to observe that the range of adopted techniques is quite wide, with models employing statistical, mathematical or even computational intelligence techniques to model AQ. More specifically, in the first category of statistical and mathematical models, a plethora of techniques, such as

the CART tool (Kaprara 2001) and the ARIMA method (Chaloulakou et al. 1999; Slini et al. 2002) have been proposed for the development of forecasting models, by associating atmospheric and meteorological data. However, the linear nature, the simplicity and the non-flexibility of these methods have, in most cases, lead to low levels of performance (low values of indexes) of the corresponding prediction models.

In the latter category of computational intelligence techniques, efficient classification models were developed by Athanasiadis et al. (2006) and Efraimidou et al. (2006), for forecasting ozone concentration levels in the urban area of Athens, using various classification methodologies. Moreover, the paper of Tzima et al. (2007) and the reported encouraging results sparked off the adoption of data mining techniques in our current work for developing AQ prediction models.

Apart from the above mentioned classification approaches, regression tasks have been extensively tackled with neural networks (NNs) in the domain of short-term AQ forecasting (Hooyberghs et al. 2005; Karatzas et al. 2008). New approaches have also been proposed, differing in the architecture of the NN (Gardner et al. 1999) or the learning algorithms employed (Nunnari 2006; Cecchetti et al. 2004; Corani 2005). In all cases, the main advantage of NN-based techniques, compared to traditional statistical methods, is their capability of approximating non-linear functions recursively in multi-scale forecasting problems. However, the training process of such techniques is quite complex and time-consuming, while their main drawback remains, regardless of the specific architecture or training algorithm used: NN-based prediction models cannot obtain real knowledge or perform physical interpretation of the underlying data sets and are, thus, not capable of generalizing into geographic areas other than the original training site.

A final family of methods that, in the last years, has drawn increasing research interest in the AQ modeling community is hybrid systems. An example of investigating the potential of such systems in AQ prediction is the work by Nunnari (2004) that combines neural and neuro-fuzzy modules in a system designed to predict episodes of poor air quality.

Moving on to the last aspect of evaluating models' performance, it is evident that experimental evaluation is an integral part of implemented research in the domain of AQ forecasting, but also in the domain of data mining in general. Despite this fact, though, it is often the case that the process of assessing experimental results is not properly handled. According to Prechlt's (1996) assessment of evaluation practices in NN learning algorithm research, one third of all articles examined do not present any quantitative comparison with a previously known algorithm. Even in the cases when a theoretical evaluation and logical analysis is applied during a

comparative study of different types of algorithms, the process – although doubtlessly significant – has a high degree of subjectivity and a high probability of incorrect or statistically invalid conclusions (Salzberg 1997).

In line with the above conclusions, our study of the referenced literature, confirms that in most cases of AQ modeling, evaluation is handled empirically, with the use of performance indexes (such as accuracy of prediction or estimated error and their significance level) and not with practical statistical methodologies. Aiming at filling this gap, our current approach employs a carefully designed evaluation methodology, combining domain-specific performance indexes with statistical tests, appropriate for comparing multiple algorithms over multiple datasets.

3 Experiments and Results

3.1 Study Area Profile

Athens is the capital city of Greece with its Metropolitan area having a population of 3.894.573 (in 2001), thus making it the largest city of the country and one of the largest urban areas in the European Union. The Athens urban area is located in a basin and it spans approximately 415 km^2. It is surrounded by high mountains (Parnitha, Pendeli, Hymettus) and fairly high hills (Aegaleo, Lycabettus, Acropolis) and it is open to the sea from the S-SW (Saronic gulf). The long-lasting warm periods and the presence of sunlight accelerate chemical reactions, thus reinforcing ozone levels and resulting in exceedances of the limits defined for pollutant concentrations.

The special topographic characteristics and meteorological conditions of the area are the most important factors influencing the levels of atmospheric pollution. The anthropogenic emissions that derive from innumerable human activities, the accumulation of industry and the high motorization are also responsible for high air pollution levels in the area. Moreover, the dense habitation and the over-concentration of population (approximately 1000 inhabitants per km^2), the poorly-designed street layout and the inconsiderable urban green space ($2m^2$ green space per citizen), are encircling emitted pollutants and increasing the time-levels of their action.

Targeting this adverse situation, our current experimental investigation involves the development of two different types of models, capable of forecasting: (i) the next hourly ozone concentrations and (ii) the 8-hour run-

ning average[1] ozone concentrations, in five locations in Athens (Agia Paraskeui-AGP, Thrakomakedones-THR, Likovrisi-LYK, Marousi-MAR, Pathsion-PAT). It is important to note, that of these two forecasting problems, the first one is formulated as a classification task – with its class attribute being nominal, while the second as a regression task – aiming at numeric predictions of the class attribute.

3.2 Data Pre-processing

The data sets used in this study, include a vast amount of air pollution concentrations and meteorological observations, recorded from the monitoring stations operated by the Hellenic Ministry for the Environment, Physical Planning and Public Works (Directorate of Air and Noise Pollution). More specifically, for each of the target stations and for all used parameters, a seven-year long data set was used, containing measurements obtained on an hourly basis, during the years 1999–2005.

In the pre-processing phase, all atmospheric and meteorological measurements were integrated in an automated way, while:

(i) additional attributes were calculated (such as the ratio of NO/O_3 and the previous 8-hour running average O_3 value, for the first and second desired forecasting models respectively),
(ii) the corresponding class attributes were defined,
(iii) missing data were handled by removing records with missing values in the class attribute, and
(iv) numeric values were transformed to nominal ones, according to relevant guidelines defining the scales presented in Table 3.1.

Table 3.1. Scales for converting pollutant concentration numerical values to nominal values

Pollutant	very low	low	medium	high	threshold	alert threshold
CO	--	0-9	10-15	16	--	--
SO_2	--	0-124	125-349	350-449	--	> 450
NO_2	--	0-199	200-249	250-359	--	> 360
O_3	0-44	45-89	90-134	135-179	180-239	> 240

The overall procedure resulted in the creation of the final datasets (one for each model type – monitoring station pair) to be utilized in the modeling process. These datasets contain from 35.636 up to 53.596 records, with

[1] According to the definition of *8-hour running average*, its value is calculated as the mean of all concentration values recorded 8-hours ahead from the current time.

15 attributes per record (including the class). More specifically, the final parameter set includes the following pollutant concentrations and meteorological measurements: (i) carbon monoxide (CO), (ii) nitrogen monoxide (NO), (iii) nitrogen dioxide (NO_2), (iv) sulfur dioxide (SO_2), (v) ozone (O_3), (vi) temperature (Ta), (vii) relative humidity (RH), (viii) wind speed (WS) and (ix) wind direction (WD). Additionally, the hour of recorded observation plus seasonal information was taken into account, such as the day of the year, the day of the week and the corresponding month. It is worth noting that no attribute selection strategy was employed; the aforementioned parameters were selected according to human experts and on the basis of their availability.

3.3 Algorithms Employed and Experimental Setup

The modeling phase was accomplished by the use of data mining algorithms, implemented in the machine learning open-source software WEKA (Waikato Environment for Knowledge Analysis) (Witten and Frank 2005). We adopted supervised learning methods, employing classification and regression data mining algorithms for the development of models forecasting next hourly ozone concentrations and 8-hour running average ozone concentrations, respectively.

Through the WEKA Explorer Environment, we experimented with all available algorithms that are grouped in 5 discrete categories, according to their functionality (Bayes, Function, Lazy, Tree and Rule-based classifiers). These categories enumerate a total of 29 classification and 15 regression algorithms, applicable to our target problems.

In order to pre-evaluate the forecasting performance of this vast number of algorithms, an intermediate step was applied: using representative data subsets for each examined area, repetitive experiments and results evaluation were performed using 10-fold cross validation, based on the indices of accuracy and correlation coefficient. As a result, 12 classification algorithms and 9 regression ones were chosen as the most efficient among the ones considered. The aforementioned pre-evaluation phase, allowed us to ensure the quality of the final results, by identifying the best performing algorithms, and to gain time and computational resources in the overall training process.

These algorithms were adopted in the next phase of developing the final AQ models (for each station – algorithm pair and for the corresponding desired type of model), where 84 different models were built using the complete training datasets and the 10-fold cross validation method.

3.4 Performance Indices

For the evaluation of the developed models' performance, we used several statistical measures and indexes of prediction accuracy, such as: the accuracy, the F-measure, the mean absolute error (MAE), the root mean squared error (RMSE) and the correlation coefficient. Additionally, we calculated two technical performance indexes, commonly reported in the domain of AQ forecasting: the Critical Success Index (CSI) and the Index of Agreement (IA). The CSI (Eq. 3.1) is a metric that combines forecasted and observed occurrences of an *atmospheric episode* without regard to successful forecasts of the non-occurred episodes, while the IA (Eq. 3.2) is an index that estimates the accuracy of a prediction model in comparison to the actual measurements.

$$CSI = \frac{A}{(A+B+C)} \tag{3.1}$$

where A: True Positives, B: False Negatives, C: False Positives

$$IA = 1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i \left(|p_i - \bar{a}| + |a_i - \bar{a}| \right)^2} \tag{3.2}$$

where p_i: predicted data, a_i: actual data, $\bar{a}$: mean actual data

It should be noted that for the models that perform next hourly ozone level predictions, an atmospheric episode is defined when the value exceeds the concentration limit of 180μg/m^3, while for the models that predict the 8-hour running average ozone levels, this limit is defined at 120μg/m^3.

3.5 Statistical Evaluation of Results

The adopted evaluation methodology was designed and implemented in multiple levels, both in terms of metrics and in sets of training parameters, so as to reinforce the validity of the final conclusions. As suggested by Demšar (2006), appropriate statistical tests were applied in each step of the methodology, to evaluate multiple algorithms over multiple data sets. We used the Friedman test to establish the significance of the differences between classifier ranks, followed by the post-hoc tests of Nemenyi and Holm to compare classifiers to each other and to verify the significance of the differences in performance among the developed predictive models.

The Friedman test (Friedman 1937) is a non-parametric method (distribution-free) that does not require homogeneity of variances. It deals with the ranked performances of k classifiers across N datasets and tests the null

hypothesis, which states that all classifiers are equivalent and, thus, their rankings R^j should be equal.

The statistical method assigns a rank $r_i{}^j$ to each of the j algorithms on all i datasets and calculates the average ranks (Eq. 3.3)

$$R^j = \frac{1}{N}\sum_i r_i^j \tag{3.3}$$

in order to define the value of the Friedman statistic (Eq. 3.4)

$$x_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{3.4}$$

which is distributed according to $x^2{}_F$ with k–1 degrees of freedom, when N and k are big enough.

Iman and Davenport (1980) showed that Friedman's $x^2{}_F$ is undesirably conservative and derived a better statistic (Eq. 3.5):

$$F_F = \frac{(N-1)x_F^2}{N(k-1)-x_F^2} \tag{3.5}$$

which is distributed according to the F-distribution with k–1 and (k–1)(N–1) degrees of freedom.

Given the calculated value of the appropriate statistic and if the Friedman test rejects the null hypothesis, we can proceed to the next step of applying the appropriate post-hoc tests. The Nemenyi test (Nemenyi 1963) is used to compare all classifiers to each other, in order to categorize them in two main groups: the ones that perform better predictions and the latter that, comparably, have significantly worse performance. For classifiers of the same group there cannot be any assessment about their statistical difference.

According to the Nemenyi test, the performance of two classifiers is significantly different if the corresponding average ranks differ by at least the *critical difference* (Eq. 3.6):

$$CD = q_\alpha \sqrt{k(k+1)/6N} \tag{3.6}$$

where critical values q_a are those of the Studentized range statistic divided by $\sqrt{2}$ with a significance level of α and k degrees of freedom (Demšar 2006). The cost of the pair-wise comparison among all classifiers is k(k–1)/2 iterations.

The Holm's step-down procedure is another post-hoc test, designed to evaluate the relative performance of the studied algorithms against a control algorithm R_O by testing hypotheses sequentially, ordered by their significance. The z statistic (Eq. 3.7), comparing the i-th classifier against the control one,

$$z = (R_i - R_0) / \sqrt{k(k+1)/6N} = (R_i - R_0) / \sqrt{SE} \quad (3.7)$$

is used to compute the corresponding probabilities from the table of normal distribution, which are, in turn, compared with an appropriate α. Given the ordered p values $p_1, p_2, \ldots, p_{k-1}$ (so that $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$), the Holm procedure starts with the most significant value and compares each p_i with the adjusted value $\alpha/(k-i)$ to compensate for multiple (k–1) comparisons. If p_i is below $\alpha/(k-i)$, the corresponding hypothesis is rejected and we proceed to comparing p_{i+1} with $\alpha/(k-i-1)$. As soon as a certain null hypothesis cannot be rejected, all the remaining hypotheses are retained as well.

The above described procedure has been applied several times with different evaluation settings: (i) per type of prediction model (next hourly O_3/ 8-hour running average O_3), (ii) per evaluation method (10-fold cross validation/ evaluation on training set), and (iii) per performance index (accuracy/ correlation coefficient/ CSI/ IA). Table 3.2 presents an extended example of a specific statistical analysis, which compares the corresponding classification algorithms according to their accuracy.

Table 3.2. Accuracy per algorithm, for next hour ozone levels prediction, using 10-fold cross validation. The ranks in parentheses are used by the Friedman test.

Classifier	AGP	LYK	MAR	THR	Average Rank
J48	81.234 (2)	83.374 (2)	84.463 (2)	85.434 (2.5)	2.125
LMT	81.286 (1)	83.346 (3)	84.519 (1)	85.431 (4.5)	2.375
NBTree	79.702 (10)	82.647 (4)	83.945 (4)	85.061 (9)	6.750
Rand.Forest	78.532 (12)	81.669 (11)	82.488 (11)	83.283 (11)	11.250
REPTree	80.097 (9)	82.348 (5)	83.159 (9)	84.934 (10)	8.250
Dec.Table	80.773 (4)	82.305 (6)	84.003 (3)	85.405 (6)	4.750
Jrip	81.114 (3)	83.383 (1)	83.669 (5)	85.239 (8)	4.250
OneR	80.334 (7)	82.122 (9)	83.264 (6)	85.436 (1)	5.750
Ridor	78.930 (11)	80.899 (12)	81.889 (12)	82.974 (12)	11.750
Logistic	80.754 (5)	81.985 (10)	83.124 (10)	85.400 (7)	8.000
Simple Log.	80.593 (6)	82.138 (7)	83.247 (8)	85.434(2.5)	5.875
SMO	80.329 (8)	82.124 (8)	83.251 (7)	85.431 (4.5)	6.875

Given the ranks in Table 3.2, the Friedman test checks whether the measured average ranks are significantly different from the mean rank R_j=6.5 expected under the null hypothesis (Eq. 3.8 and Eq. 3.9):

$$x_F^2 = \frac{12 \cdot 4}{12 \cdot 13}[(2.125^2 + 2.375^2 + 6.75^2 + 11.25^2 + 8.25^2 + 4.75^2 + 4.25^2 + 5.75^2 + 11.75^2 + 8^2 + 5.875^2 + 6.875^2) - \frac{12 \cdot 13^2}{4}] = 31.038 \quad (3.8)$$

$$F_F = \frac{3 \cdot 31.038}{4 \cdot 11 - 31.038} = 7.184 \tag{3.9}$$

With 12 algorithms and 4 datasets, F_F is distributed according to the F-distribution with 11 and 33 degrees of freedom. The critical value of $F(11,33)$ for α=0.05 is 2.093, so we reject the null hypothesis.

Further analysis on the same case involves the application of the Nemenyi and Holm post-hoc tests. The critical value q_a is 3.268 at α=0.05 and 3.030 at α=0.10. Therefore the CD is 8.332 and 7.725 respectively, which means that difference in average ranks of algorithms greater than the value of CD is significant.

The statistical differences between the algorithms, obtained from the Nemenyi test, are visually represented with a simple CD-diagram (Demšar 2006), as shown in Fig. 1. Classifiers are lined up in the diagram, according to their average ranking (see last column in Table 3.2), and those who are connected with a straight bold line (defines the CD) are not significantly different with each other. A careful examination of the diagram reveals that J48 and LMT are significantly better than Random Forest and Ridor, since they are not connected with the straight bold line.

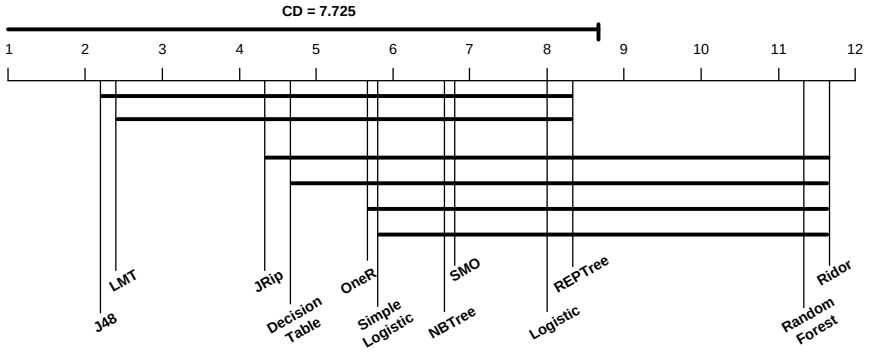

Fig. 1 Comparison of all classifiers in Table 3.2 against each other, using the Nemenyi post-hoc test.

The evaluation methodology as described above (including the application of the Holm post-hoc test) was applied on all algorithm – station pairs for both our target prediction tasks. This analysis provided us with valuable insights on the potential of the studied algorithms in the task of predicting AQ. Overall, the developed models exhibit high percentages of correct predictions (82-84%) in the set of target monitoring stations and satisfactory values for domain-specific calculated metrics (CSI [O_3]: 0.4, CSI [8-hour av.O_3]: 0.6, IA: 0.92-0.96), as summarized in Table 3.3.

More importantly, the results obtained from the thorough statistical analysis applied on all developed models, revealed that in the task of forecasting AQ in Athens:

(i) J48, LMT, JRip, Decision Table and REPTree (belonging to the categories of "Tree" and "Rule-based classifiers") are the best performing classification algorithms, while

(ii) M5P, REPTree and M5Rules (belonging to the same categories as the above) are the best performing regression algorithms.

Table 3.3. Average performance of classification and regression models, for next hourly ozone levels and 8-hour running average ozone levels prediction respectively.

Classification algorithms	**% Correctly Classified**	**MAE**	**CSI**	**IA**
J48	83.626	0.091	0.396	0.935
LMT	83.646	0.086	0.419	0.935
JRip	83.351	0.091	0.380	0.925
Dec.Table	83.121	0.086	0.396	0.930
REPTree	82.634	0.086	0.395	0.931
Logistic	82.816	0.092	0.402	0.932
Regression algorithms	**Correlation Coefficient**	**MAE**	**CSI**	**IA**
M5P	0.897	9.163	0.575	0.953
REPTree	0.895	8.936	0.658	0.967
M5Rules	0.887	9.681	0.552	0.944

On the other hand, prediction models based on SMOreg and Linear Regression (from the "Function classifiers" category) perform statistically worse. Finally, Random Forest is proved to be weak in cases of generalized evaluation with supplied test sets from different monitoring stations.

4 Conclusions and Future Work

In this paper we have adopted and evaluated DM methods for studying the AQ modeling domain. We investigated the levels of their performance in real data and proposed an integrated statistical methodology for a well-founded evaluation process of the quantitative results. On the basis of the reported findings, we consider our investigation successful, in terms of our original goals of: (i) discovering the best performing set of algorithms for the development of AQ prediction models and (ii) employing a carefully

designed statistical methodology in the evaluation process of the developed AQ forecasting models.

Of course, several issues deserve further investigation, including the utilization of additional environmental parameters and the study of existent periodicity among the values of atmospheric/ meteorological data. Another important aspect is the study of probable interrelations between neighboring locations and the algorithms' potential in the generalization process, through the evaluation of the developed models with unknown data.

Acknowledgments

The second author acknowledges that this paper is part of the 03ED735 research project, implemented within the framework of the "Reinforcement Programme of Human Research Manpower" (PENED) and co-financed by National and Community Funds (25% from the Greek Ministry of Development – General Secretariat of Research and Technology and 75% from E.U. – European Social Funding).

References

Athanasiadis IN, Karatzas K & Mitkas PA (2006) Classification techniques for air quality forecasting. Proceeding of the 5th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, Riva del Garda, Italy

Cecchetti M, Corani G & Guariso G (2004) Artificial Neural Networks Prediction of PM10 in the Milan area. 2nd International Environmental Modelling and Software Society Conference, Osnabruck

Chaloulakou A, Assimacopoulos D & Lekkas T (1999) Forecasting Daily Maximum Ozone Concentrations in the Athens Basin. Environmental Monitoring and Assessment, vol 56, no 1, pp 97-112

Corani G (2005) Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling 185, pp 513-529

Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, vol 7, pp 1-30

Efraimidou M, Kanaki M, Athanasiadis IN, Mitkas P & Karatzas K (2006) Data mining air quality data for Athens, Greece. Proceeding of the 20th International Conference on Informatics for Environmental Protection, Graz, Austria, pp 505-508

Friedman M (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association, vol 32, pp 675-701

Gardner MW & Dorling SR (1999) Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment, vol 33, no 5, pp 709-719

Goebel M & Gruenwald L (1999) A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations Newsletter, vol 1, no 1, pp 20-33

Hooyberghs J, Mensink C, Dumont G, Fierens F & Brasseur O (2005) A neural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment, vol 39, no 18, pp 3279-3289

Iman RL & Davenport JM (1980) Approximations of the critical region of the Friedman statistic. Communications in Statistics, vol 9, issue 6, pp 571-595

Kaprara A, Karatzas K & Moussiopoulos N (2001) Maximum Ozone level prediction in Athens with the aid of the CART system, a modelling study. Proceedings of the VII International Conference on Harmonization within Atmospheric Dispersion Modelling for Regulatory Purposes, Beligerate (Lake Maggiore), Italy, pp 193-196

Karatzas K, Papadourakis G & Kyriakidis I (2008) Understanding and forecasting atmospheric quality parameters with the aid of ANNs. IEEE World Congress on Computational Intelligence

Nemenyi BP (1963) Distribution-free multiple comparisons. PhD thesis, Princeton University

Nunnari G (2006) An improved back propagation algorithm to predict episodes of poor air quality. Soft Computing - A Fusion of Foundations, Methodologies and Applications, vol 10, no 2, pp 132-139

Nunnari G, Dorling S, Schlink U, Cawley G, Foxal R & Chatterton T (2004) Modelling SO2 concentration at a point with statistical approaches. Environmental Modelling & Software, vol 19, no 10, pp 887-905

Prechelt L (1996) A Quantitative Study of Experimental Evaluations of Neural Network Learning Algorithms: Current Research Practice. Neural Networks, vol 9, issue 3, pp 457-462

Salzberg LS (1997) On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery, pp 317-327

Slini T, Karatzas K & Moussiopoulos N (2002) Statistical analysis of environmental data as the basis of forecasting: an air quality application. The Science of the Total Environment, vol 288, pp 227-237

Tzima FA, Karatzas KD, Mitkas PA & Karathanasis S (2007) Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece. International Joint Conference on Neural Networks, pp 2752-2757

Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques (2nd Edition). Morgan Kaufmann, San Francisco, CA, USA