In **Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence**, Riva del Garda, Italy, August 2006.

Classification techniques for air quality forecasting

Ioannis N. Athanasiadis[1] and **Kostas D. Karatzas**[2] and **Pericles A. Mitkas**[3]

Abstract. Air quality forecasting is one of the core elements of contemporary Urban Air Quality Management and Information Systems. Such systems are usually set up in order to serve environmental legislation needs and are tailored towards decision makers (for atmospheric quality problem abatement) and citizens (for early warning and information provision). The pluralism of forecasting methods that are available does not always lead to forecasting success, as the specific characteristics of each area of interest and the complicated, mostly chaotic relationships between air quality, meteorology, emissions and topography, limit the effectiveness of the methods used. On the other hand, the timescale of air quality problems dictate the usage of relatively "fast" methods, while the varying quality of input data calls for methods that have a low sensitivity in this factor and a high operational potential. For this reason, it is always interesting to perform a comparative study between various air quality forecasting methods and tools. The present paper describes the comparison work performed between several statistical methods and classification algorithms, on the basis of their performance for specific air quality time series in Athens, Greece.

1. INTRODUCTION

Urban air quality information is created when methods, tools or human judgment is applied over a data set that is usually comprised of time series records resulting from the operation of monitoring stations. Mathematical methods and tools may provide with forecasting capabilities, thus offering decision makers with the opportunity to take preventive measures that would "smooth" or alter the results of a forecasted "episode" or even "crisis". The complexity of air pollution data has been extensively discussed [1], while the usage of various modelling tools is frequently addressed in related literature [2]. Earlier research work has dealt with using knowledge discovery techniques mainly for air quality associated incident forecasting, on the basis of incident definitions (i.e. episodes, pollution threshold value exceedances etc) that are associated with air pollution levels. One should make a distinction between (a) the usage of three-dimensional air pollution models that are used for the assessment of air pollution levels and for the forecasting of air quality on a grid basis concerning an area of interest, and (b) the usage of mathematical methods that deal directly with time series "generated" by monitoring stations. For the latter, several models have been built for predicting incidents that may occur in the near future. For instance, conventional statistical regression models [3], [4], [5], and time-series analysis have been applied to predict ozone levels [6]. Neural networks have been used for short-term ozone prediction [7], also in combination with adaptive nonlinear state space-based prediction [8]. In addition, methods like case-based reasoning [9] and classification and regression trees [10], [11], have been employed for predicting air pollutant concentrations. In the present paper several classification algorithms are applied, for analyzing air quality information and for forecasting ozone concentration levels in a dense urban area. Results obtained are compared with those coming from the application of various statistical methods, and specifically (a) Auto-Regressive Integrated Moving Average model (ARIMA), (b) Linear Regression Analysis (LRA), and (c) Principal Components Analysis (PCA), that have been published previously for the same data set [12]. In addition, more detailed experiments are performed for the identification of forecasting performance of various classification algorithms.

[1] Dalle Molle Institute for Artificial Intelligence, Galleria 2, CH6908 Lugano, Switzerland, ioannis@idsia.ch

[2] Dept. of Mechanical Engineering, Aristotle University, GR 541 24 Thessaloniki, Greece, kkara@eng.auth.gr

[3] Dept. of Electrical and Computer Engineering, Aristotle University, GR541 24 Thessaloniki, Greece, mitkas@eng.auth.gr

2. AIR QUALITY IN ATHENS

2.1 Historical aspects

One of the first ever recorder efforts in dealing with air pollution problems in Athens is related to the fundamental legislative works of Solon in the 6th century BC, where, among others, he ruled that blacksmiths activities should be transferred outside the city of Athens in order to avoid noise and air pollution. In recent times, air quality problems were firstly addressed in 1963, where the analogy between Los Angeles and Athens was introduced in relation to the effects of traffic related air emissions to the quality of breathing air [13]. Later on, a first attempt for a qualitative assessment of air pollution levels in the city was made [14], while Zampakas [15] reports on sodium dioxide (SO_2), black smoke and 3,4 venzopyrene levels in the area of 391, 700 and 62 μg/m^3 respectively [15]. At the same time, the relationship between prevailing meteorological conditions and air pollution levels is firstly identified. Concerning meteorology, it should be noted that one of the most interesting features of air flow patterns in the Greater Athens Area is attributed to the development of sea and land breeze circulation systems, that, astonishingly enough, seem to prevail in the area continuously, for the last 3000 years at least. More specifically, according to Thucydides [16] in the battle between the Athenian and Peloponnesian in the Gulf of Corinth, 429 BC, Athenians delayed the start of their attack until a morning wind "from the Gulf" of Corinth would turn the sea choppy and make the rowing hard to inexperienced rowers. In conclusion Thucydides' brief account of the wind and battle in the Gulf of Patras can only be reconciled with the knowledge of wind conditions if we assume that his description is not necessarily

accurate as far as details of "precise" direction and time of rise of speed of the wind. In Aeschyli: Persae, (472 BC) the great tragic poet, describes (from own experience) the ship battle between the Athenians and the Persians in Salamina (480 BC) where the autumn sea breeze played a role in the ability of battleships to perform manoeuvres [17].

2.2 The city of Athens

The city of Athens today is located in a basin of approximately 450 km^2 (Figure 1). This basin is surrounded at three sides by fairly high mountains (Mt. Parnis, Mt. Pendeli, Mt. Hymettus and Mt. Aegaleon), while to the SW it is open to the sea. Industrial activities take place both in the Athens basin and in the neighbouring Thriasion plain. The Athens basin is characterized by a high concentration of population (about 40% of the Greek population), accumulation of industry (about 50% of the Greek industrial activities) and high motorization (about 50% of the registered Greek cars). Anthropogenic emissions in conjunction with unfavourable topographical and meteorological conditions are responsible for the high air pollution levels in the area.

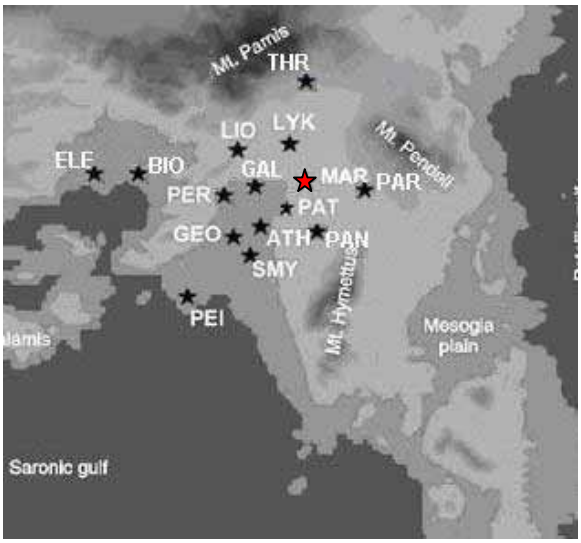

Figure 1. Topography of the Greater Athens Area and location of the Air Quality monitoring stations. Marousi station is indicated with red.

The visual results of atmospheric pollution “nephos”, a brown cloud over the city, made their appearance in the 70’s. Alarmingly elevated pollutant concentrations already threaten public health and at the same time cause irreparable damage to invaluable ancient monuments. Already by mid 80’s it became apparent to the Athenian authorities that road traffic is the main contributor to the anthropogenic emissions in Athens: Practically all carbon monoxide (CO) emissions, 3/4 of the nitrogen oxides (NO_x) emissions and nearly 2/3 of the volatile organic compounds (VOCs) emissions were found to be associated with road traffic. It was realized that the largest part of these emissions was associated with private use passenger cars. Therefore, decisive interventions to the market and use of cars were recognized as essential for improving air quality in Athens.

2.3 Air quality monitoring in Athens

The existing network in Athens comprises today 17 fully automatic monitoring stations for which the Hellenic Ministry of Environment, Physical Planning and Public Works (HEMINENV, http://www.minenv.gr) is responsible (data being available since 1983). Station locations are provided in Figure 1: location PEI hosts two stations, while north of ATH and very close to it (thus not visible), is another station, named Aristotelous. The station of interest for the current paper is MAR, i.e. Marousi, an urban station in the north part of the city, influenced by sea-land breeze circulations and under the influence of traffic emissions. Stations monitor air pollutants as well as meteorological parameters and are connected with a central automatic collection and processing unit, located in one of the Ministry of the Environment buildings near the centre of Athens. The monitoring network is operated, managed and maintained exclusively by HEMINENV (Directorate of Atmospheric Pollution and Noise Control/Department of Atmospheric Quality). All data collection, processing and maintenance operations is done by HEMINENV officials and technical personnel. HEMINENV is responsible for issuing short-term emergency bulletins on days of high pollution levels and imposing various restrictive measures to limit the magnitude of pollution incidents. Long-term strategic policy planning and implementation of permanent measures is also a HEMINENV responsibility.

3. AIR QUALITY MANAGEMENT AND OZONE

The assessment and management of air quality (AQ), especially in urban areas, is characterised by major uncertainties that are vis-à-vis characterised by, and extend well beyond, the boundaries of the atmospheric chemistry and physics and the stochastic nature of the major air pollutant emission mechanisms (including those of anthropogenic origin). Air pollution is a problem that can not be treated independently of the urban web. One has to consider the urban environment as a multi-dimensional, multivariable system, part of which is the AQ aspect, that also includes the layout of the city, the existence of green and un-built areas, the geometry, the architectural morphology and the thermal properties of the buildings, the vehicular traffic, the stationary thermal systems and of course the local microclimatic conditions. Concerning photochemical pollutants, it should be noted that their dynamic nature, accompanied by the strong non-linearities in the underlying physical and chemical mechanisms involved in their creation, chemical transformation and transportation-diffusion, was always among the major challenges for the development of any modelling – forecasting method and tool.

Focusing on ozone (O_3), one has to take into account that it is a secondary pollutant and is formed as a result of reactions between pollutants emitted from industrial sources and automobiles. The precursors of ozone are nitrogen oxides and volatile organic compounds. In the presence of sunlight (ultraviolet radiation) and, under suitable meteorological conditions, the precursors react photochemically to “produce” ozone. Ambient concentrations of ozone depend on weather related conditions, because the amount of reactants, the reaction rates and atmospheric dispersion are sensitive to meteorological conditions changes [18].

4. METHODOLOGY

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends [19]. Classification predicts categorical labels, the co-called "classes", while prediction models are used for forecasting continuous valued variables. Typically, both classification and prediction analyses are a two-step process. In the first step, a classification/prediction algorithm is applied on the available data and a decision model is extracted. The mined decision model encapsulates the knowledge lying in the data in a form such as a decision tree, a neural network, a regression model, a support vector machine, etc. This step is usually called model training phase. The second step is the testing phase, which involves the application of the decision model on data for making decisions (predictions). This phase focuses on testing the ability of the decision model to approximate data not used in training. In this respect, both classification and statistical prediction algorithms have been applied for function approximation in several, yet diverse, domains. To distinguish the terms "classification" and "prediction", the reader may have in mind the following working definition: *Whereas classification predicts qualitative variables, prediction forecasts quantitative ones.*

As it has already been mentioned, in the air quality domain, several data-driven forecasting models have been developed, using a variety of prediction methods. These models should be considered separately from deterministic models, which apply a large set of physical and chemical laws trying to “clone” the real environment into a simulated one, and thus forecast the parameters of interest. In addition, data driven models should also be distinguished by models associated with the treatment of both numerical (deterministic) models and observation data, like data assimilation models. A review and a classification of air quality forecasting models is available by [20] and [21], while the importance of “fast” air quality forecasting methods is highlighted, among others, in [22].

In all data-driven approaches air quality data has been analyzed following a "prediction" perspective, i.e. the models developed are approximating a continuous (quantitative) variable (a pollutant's concentration). Such models are sufficient for administrative operations, scientific tasks and the study of the environmental phenomena associated with air quality degradation. However, such an approach is unsuitable for using data-driven models in an operational manner. Air quality assessment, from an operational point of view, requires the characterization of atmospheric quality using qualitative indicators on the basis of a quantitative classification. Legislative acts, as the US Clean Air Act and the 96/62/EU framework directive for urban air management (and the accompanying daughter directives) have delimited certain thresholds for characterizing the quality for the atmospheric environment. In this background, qualitative information (as the air quality indicators) is identified as being of great importance for any operational Air Quality Management System. Motivated by this remark, this paper presents a comparative study between statistical methods and classification algorithms for contemporary air quality forecasting (Figure 3). For this purpose, a set of environmental data have been used, consisting of time series information that include ozone concentration level observations available for a 3.5-year long time period (January 1999 - June 2002), for Athens, Greece (Figure 2). The present study is focused on an urban monitoring station at the north-eastern suburbs of the city (Marousi).

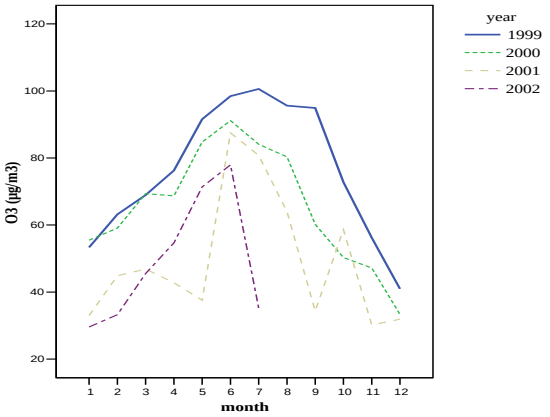

Figure 2.Evolution of the monthly mean O_3 concentrations for Marousi.

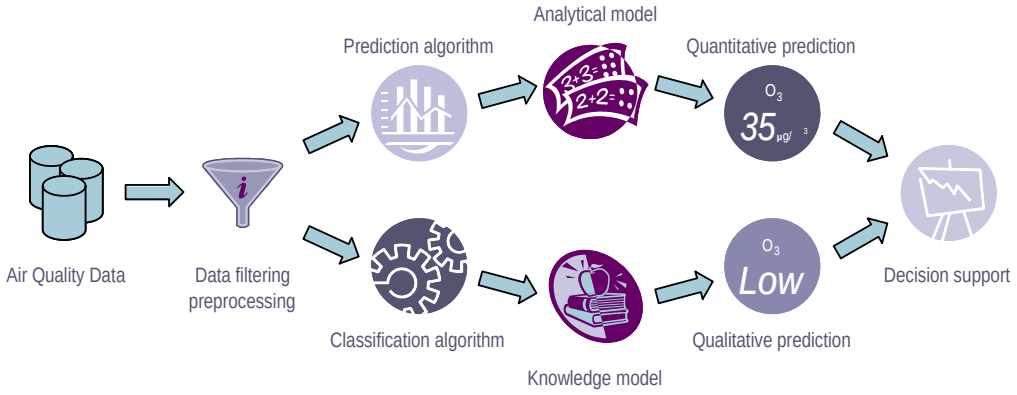

Figure 3. The classical statistical regression algorithms and the classification algorithms approach (upper and lower leg respectively)

5. METHOD APPLICATION, RESULTS AND DISCUSSION

As air quality information is by definition stochastic, chaotic and difficult to inter-relate, data mining offers automatic rule extraction as a feasible solution: by searching through large amounts of data, is possible to identify sufficient instances of an association between data value occurrences to suggest a statistically significant rule. However, a domain expert is still required for effectively map the domain knowledge to extraction rules and to action sequences. In addition to that, data have to be "pre-processed", i.e. cleaned (for errors or gaps), explored (for acquaintance with the real world parameters they represent) and transformed (in case that such an action is required for data homogenisation purposes).

On this basis, and aiming at investigating the operational performance of state-of-the-art data mining methods and tools towards ozone forecasting in urban areas, a number of classification-based methods were employed. In contrary with conventional statistical approaches, classification algorithms used in this paper utilize other criteria/functions, such as the information gain and entropy, distance-metrics, or memory-based learning, for encapsulating data-driven knowledge and ultimately drawing conclusions. In this work, different types of algorithms for classification are evaluated. Among them, instance-based learners are used, as iBK, Kstar, Nnge (Nearest Neighbor With Generalization), rule-based classifiers, as Conjunctive Rules, OneR, Decision Tables, decision trees (C4.5, ADTrees), along with Bayesian Classifiers (NaiveBayes), and Neural networks (Voted Perceptron). All these algorithms are implemented in WEKA [23]. The WEKA platform, which is open source software, was used for the data mining experiments described below. For further information on the algorithms, the reader may refer to [24].

5.1 Experiments for the daily vegetation threshold

Two series of experiments have been conducted for the available data from the Marousi station. Ozone concentration has been used in both cases as the basis for assessing urban air quality. The first experiment has been formulated as a two-class categorization problem, so that to be directly comparable with the results obtained using statistical regression and reported in [12]. As an early warning threshold was set to be the daily alert threshold the limit of 65 μg/m^3, which is the vegetation protection threshold for the mean value over 24 hours according to the former European Directive 92/72/EEC. The focus was the successful forecast of daily mean ozone concentration values, through the application of the vegetation protection threshold for the mean value over 24 hours according to the former European Directive 92/72/EEC.

The available data has been split in two parts: one for training (that corresponds to the period January 1999 – December 2001) and one for testing (January – June 2002). The purpose of this selection was the ability to make comparisons with the statistical methods used for the same test case as previously reported.

For LRA the ozone lagged observation values (i.e. ozone concentration level of the previous day), nitrogen oxides (NOx) concentration level, minimum temperature and the week factor were included in the model. For ARIMA, the model ARIMA(0,1,1) including the mean wind speed best fitted the data, while in the case of PCA, 4 principal components were created accounting for temperature, wind speed, ozone precursors (NOx, NO) and time variables respectively. For all classification models evaluated, ozone precursors, other pollutant concentrations and meteorological data were used as model inputs.

Results obtained on the testing phase for mean daily ozone concentration levels are summarised in Table 1, where, in parallel, results obtained for the same test case with statistical methods (LRA, ARIMA and PCA) and published in [12], are included for comparison purposes. For each model, the root mean square error (RMSE) which summarizes the difference between the observed and modeled concentrations, and the percentage of the correct and false alarms are given, in combination with the so called Kappa Index (i.e, the Critical Success Index). It should be noted that the Kappa Index is calculated on the basis of the occurrence of events (successful/unsuccessful forecasts in combination of occurrences of threshold value exheedances), on the basis of Table 2, [25] and by making use of eq. (1)

Table 2. Contingency table (Confusion matrix) for threshold forecasts

Exceedances	Forecasted	
Observed	Yes	No
Yes	a	b
No	c	d

$$\text{Kappa Index} = \frac{a}{a+b+c} \quad (1)$$

Table 1. Statistical performance of data mining models accompanied by statistical models, for Marousi monitoring station. Test case: mean daily O_3 concentration levels, with threshold set to 65 μg/m^3.

	Model	RMSE	Correct alarms (%)	false alarms (%)	Kappa Index
Statistical regression algorithms	ARIMA	0.153	31.58	2.94	0.154
	LCA	0.128	53.37	0.37	0.256
	PCA	0.261	21.47	0.55	0.148
Classification algorithms	NaiveBayes	0.404	80.65	17.65	0.154
	Voted Perceptron	0.414	80.65	15.97	0.630
	IBk	0.438	59.68	15.13	0.458
	KStar	0.598	61.67	43.70	0.160
	Conjunctive Rule	0.398	64.52	10.92	0.555
	Decision Table	0.381	69.35	12.61	0.576
	Nnge	0.44	59.68	8.40	0.544
	OneR	0.42	70.97	11.76	0.601
	ADTree	*0.361*	*85.48*	*17.65*	*0.648*
	C4.5 (J48)	0.376	75.81	19.33	0.548

One first overall comment is related to the differences concerning the RMSE values concerning the classification methods (ranging from 0.376 to 0.598) and the statistical methods (ranging from 0.128 to 0.261). Although this may be considered to be a criterion for poor modelling performance, it should be noted that statistical methods are based on the forecast of a continuous set of values, while classification methods are based on the forecast of occurrences that are either true or false; thus, RMSE is not a sufficient criterion for model performance concerning classification methods, and should be used solely to demonstrate the scattering of values generated by predictions in comparison to observed ones. Coming to the percentage of correct alarms forecasted, it is evident that classification models perform better compared to statistical models, with performance ranging from 59.68% (IBk - K-nearest neighbours classifier) to 85.38% (ADTree-Alternative Decision Trees), while for the percentage of wrong alarms, classification models perform between 8.4% and 43.7%. Overall, with the aid of the Kappa Index, it is evident that classification models have a much better performance for the specific test case in comparison to statistical models, as their success index rages from 0.154 to 0.648, while for statistical models does not exceed 0.26. On the basis of obtained results, the classification algorithms seem to have an advantage in comparison to statistical ones, achieving better performance concerning air quality management-related decisions taken on the basis of threshold values used. Yet, more analysis and testing should be made in order to arrive to a sufficient proposal for an effective ozone forecasting (OF) module for an urban area of interest. Thus, the combined usage of classification methods and statistical methods may provide an effective OF module for operational usage that will enhance environmental decision making at an urban level.

5.2 Experiments for multiple hourly ozone value range categories

In order to further explore the performance of the various classification algorithms in air quality forecasting, hourly ozone values were investigated for the same station (Figure 4).

The goal of this experiment was to investigate the performance of air quality forecasting models concerning the maximum hourly ozone concentration value for the next 8, 24, 48 and 72 hours. For this reason, the following categorization was used, similar to the approaches used in various EU countries (Directive 2002/3/EU and http://www.luftkvalitet.info/)

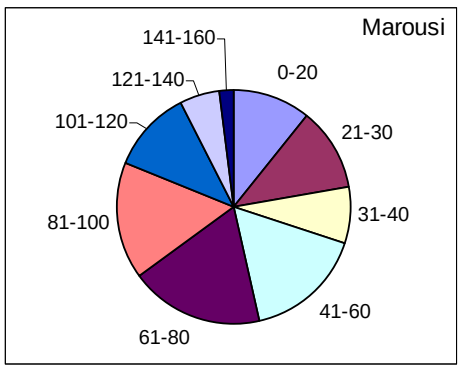

Figure 4. Distribution of hourly ozone concentration values (in μg/m^3) for the time period studied (1/1/1999 – 1/7/ 2002).

Table 3. Ozone concentration categorization

Value range (in μg/m^3)	Class	Explanation
0-44	vl	very low
45-89	l	low
90-134	m	medium
135-179	h	high
180-239	i	information threshold
>240	a	alert threshold

In addition, and in order to evaluate forecasting performance, "costs" were empirically introduced and were associated with false forecasts, as presented in Table 4 below.

Table 4. Costs for false forecasts

classified as →	**vl**	**l**	**m**	**h**	**i**	**a**
vl	0.0	0.4	0.8	1.5	3.0	5.0
l	0.4	0.0	0.4	1.0	2.5	4.0
m	0.8	0.4	0.0	0.6	2.0	3.5
h	1.5	1.0	0.6	0.0	1.5	3.0
i	3.0	2.5	2.0	1.5	0.0	2.0
a	5.0	4.0	3.5	3.0	2.0	0.0

Thus, when a value that actually belongs to class *l* is falsely categorized as belonging to class m, the error caused has a "cost" of 0.4, while if a value belongs to class *a* but is reported as belonging to class *i*, the cost rises to 2.0, as in the later case the difference between the two classes is the occurrence of the alerts, which is considered to be a much more severe situation. On this basis, a set of classification algorithms were evaluated: BayesNet, NaiveBayes, Multilayer Perceptron (MLP), IBk, Decisionstump, C4.5(J48), Conjunctive Rule, Nnge (Nearest Neighbour With Generalization) and OneR (1R). The application of the aforementioned algorithms was done with some variations concerning parameters used (not discussed in the present paper), and taking into account meteorological information. The performance accuracy for the testing phase is presented in Table 5.

The overall forecasting performance of all algorithms is within accepted range. The most efficient algorithm for the experiments conducted is C4.5(J48), which is a decision tree learning algorithm. Results performance is high for all time-frames studied, suggesting that classification methods may be considered for the operational provision of air quality forecasts to citizens and decision makers. It should be noted that the robustness of predictions is a function of the "arbitrary" definition of cost factors, yet the latter allow for prioritizing forecasting methods in relation to their relevant performance, concerning the case studied here.

Table 5. Percentage of correct alarms concerning maximum hourly ozone concentration value forecast for the next 8, 24, 48 and 72 hours for the Marousi monitoring station.

	Time-frames (hours in advance)			
	8 h	24 h	48 h	72 h
OneR	61.49	71.82	67.09	62.90
Nnge	75.34	86.85	87.89	87.94
Nnge default	73.83	86.94	88.47	88.09
Conjunctiverule	60.39	63.46	56.13	47.36
C4.5 (J48)	*78.09*	*92.32*	*95.64*	*96.22*
DecisionStump	59.67	63.43	55.61	49.03
Ibk	55.92	56.04	54.78	55.72
NaiveBayes	63.35	70.83	68.97	66.21
BayesNet	61.36	70.61	65.61	64.32
MLP 3	70.47	72.69	66.96	64.48

6. DISCUSSION

The primary goal of the study was the application of classification data mining algorithms for the development of an operationally efficient OF module and the comparison of their performance with statistical analysis methods. For this reason, a set of air quality data from the Marousi air quality monitoring station, in Athens, Greece, was used. The comparative analysis of the models performance showed that for the specific test case the classification algorithms have a considerably better performance compared to the statistical methods. Also, it has been discussed elsewhere [26,27] that the understainability of the decision models generated by some of the classification algorithms could have a significant impact on the trustworthiness of the models. In addiiton, empirical results achieved in this work, suggest that classification methods could be employed together with statistical methods and other "fast" data analysis and prediction algorithms, for the creation of operational air quality forecasting modules, that may effectively support operational air quality management on a day-to-day basis, in line with contemporary EU environmental legislation. These findings were supported by a second set of experiments, aiming at the forecasting of maximum hourly ozone concentration levels for the next 8, 24, 48 and 72 hours, with data coming from the same station and for the same time period. The performance achieved concerning the percentage of correct alarms suggests that classification methods should be considered as appropriate for operational air quality forecasting applications. Given the encouraging results of this study, future efforts will concentrate on combining simple classification models into aggregated classifiers, (i.e. ensemble learning).

REFERENCES

[1] F.M. Morabito and M. Versaci, 'Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data', *Neural Networks*, **16,** 493–506, (2003)

[2] M. Makowski, 'Modeling paradigms applied to the analysis of European air quality, *European Journal of Operational Research*, **122**, 219-241, (2000).

[3] S. Bordignon, C. Gaetan and F. Lisi, 'Nonlinear models for ground-level ozone forecasting', *Statistical Methods and Applications*, **11**, 227-246, (2002).

[4] L. S. Huang. and R. L. Smith, 'Meteorologically-dependent trends in urban ozone' Technical Report 72, National Institute of Statistical Sciences (1997).

[5] Kim H.Y. and Guldmann J.M. 'Modeling air quality in urban areas: A cell-based statistical approach', *Geographical Analysis*, **33,** (2001).

[6] Chen L.-J., Islam S. Biswas P. 'Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction', *Atmospheric Environment*, **32**, 1839-1848, (1998).

[7] S.M. Shiva Nagendra and M. Khare, 'Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions', *Ecological Modelling*, **190,** 99–115, (2006).

[8] A. Zolghadri,, M. Monsion, D. Henry, C. Marchionini, and O. Petrique, 'Development of an operational model-based warning system for tropospheric ozone concentrations in Bordeaux, France', *Environmental Modelling & Software*, **19,** 369–382, (2004).

[9] E. Kalapanidas and N. Avouris, 'Short-term air quality prediction using a case-based classifier', *Environmental Modelling & Software*, **16**, 263–272, (2001).

[10] M.A. Barrero, J.O. Grimalt, L. Canto´n, 'Prediction of daily ozone concentration maxima in the urban atmosphere', *Chemometrics and Intelligent Laboratory Systems*, **80**, 67– 76, (2006).

[11] D. J. Briggsa,, C. de Hoogh, J. Gulliver, J Wills, P. Elliott, S. Kingham and K. Smallboned, 'A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments', *The Science of the Total Environment*, **253,** 151-167, (2000).

[12] Th. Slini, K. Karatzas and N. Moussiopoulos, 'Ozone forecasting supported by data mining statistical methods, *Proceedings of the 5th International Conference on Urban Air Quality Measurement, Modelling and Management*, (R. Sokhi and J. Brexhler eds), Valencia, 29-31 March 2005, Spain, ISBN I-898543-92-5.

[13] A. Mukhopadhyay, 'Air pollution by automotive source in urban centres with reference to Athens metropolitan area', MSc thesis, Athens, 25 of August 1963'. Athens Technological Institute Graduate School of Ekistics. ACE Publication Series Research Report No 8, 1970.

[14] J. Papaioannou, 'Air pollution in Athens', *Ecistics*, **25**, 72-80, (1967).

[15] I. Zampakas, 'Meteorological conditions for minimum and maximum air pollution levels in Athens', Laboratory of Climatology, University of Athens (in Greek, 1973).

[16] J. Neumann and D.A. Metaxas, 'The battle between the Athenian and Peloponnesian fleets, 429 BC and Thucydides, wind from the Gulf (of Corinth)', *Meteorologische Rundshau*; **32,** 182-188 (1979).

[17] K. Karatzas, 'Preservation of environmental characteristics as witnessed in classic and modern literature: the case of Greece, *The Science of the Total Environment*, **257**, 213-218 (2000).

[18] MacDonald C., Roberts P., Main H., Dye T., Coe D., Yarbrough J., 'The 1996 Paso del Norte Ozone Study: analysis of meteorological and air quality data that influence local ozone concentrations', *The Science of The Total Environment* **276**, 93-109 (2001).

[19] Han J. and Kamber M. Data Mining: Concepts and techniques Morgan Kaufmann Publishers (2001).

[20] M. Copeand D. Hess 'Air quality forecasting: A review and comparison of the approaches used internationally and in Australia', *Clean Air and Environmental Quality*, **39**, 52-60 (2005), http://www.casanz.org.au/techpapers.htm.

[21] Wayland, R. A., J. E. White, P. G. Dickerson, and T.Dye, 2002: Communicating real-time and forecasted air quality to the public: current state and future plans, *Environmental Management*, 28-36.

[22] R. Husar and R. Poirot, 'DataFed and Fastnet: Tools for Agile Air Quality Analysis' *Environmental Manager* 2005, September, 39-41, available via http://capita.wustl.edu/capita/capitareports/050601AWMA_FASTNET/Submitted/EM_DataFed_FASTNET_050720.pdf

[23] WEKA, The Waikato Environment for Knowledge Analysis, www.cs.waikato.ac.nz/ml/weka, Version 3.4, 2004

[24] Witten I. and Frank E. 1999, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.

[25] McHenry J., McKeen S., Ryan W., Seaman N., Pudykiewicz J., Grell G., Stein A., Coats Ch. and Vukovich J., 2003, Forecasting air quality with MODELS-3 components: performace expectations, Models-3 User's Workshop http://www.cmascenter.org/2003_workshop/downloads.html

[26] Athanasiadis, I.N. and Mitkas, P.A. (2004), Supporting the decision-making process in environmental monitoring systems with knowledge discovery techniques, in H. Voss; M. Wachowicz; S. Dzeroski & A. Lanza (eds).,'Knowledge-based Services for the Public Sector Symposium', vol. III: Knowledge Discovery for Environmental Management', KDnet, Bonn, Germany, pp. 1-12.

[27] Athanasiadis, I. N. and Kaburlasos, V.G., Air quality assessment using Fuzzy Lattice Reasoning (FLR), to appear in 2006 IEEE International Conference on Fuzzy Systems, Vancouver, Canada .