Expert Systems with Applications 38 (2011) 5019–5027

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Sparse episode identification in environmental datasets: The case of air quality assessment

Fani A. Tzima [a,*], Pericles A. Mitkas [a], Dimitris Voukantsis [b], Kostas Karatzas [b]

[a] *Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki GR-541 24, Greece*
[b] *Department of Mechanical Engineering, Aristotle University of Thessaloniki, Thessaloniki GR-541 24, Greece*

ARTICLE INFO

Keywords:
Air quality (AQ)
Sparse episode identification
Domain-driven data mining
Model evaluation

ABSTRACT

Sparse episode identification in environmental datasets is not only a multi-faceted and computationally challenging problem for machine learning algorithms, but also a difficult task for human-decision makers: the strict regulatory framework, in combination with the public demand for better information services, poses the need for robust, efficient and, more importantly, understandable forecasting models. Additionally, these models need to provide decision-makers with "summarized" and valuable knowledge, that has to be subjected to a thorough evaluation procedure, easily translated to services and/or actions in actual decision making situations, and integratable with existing Environmental Management Systems (EMSs).

On this basis, our current study investigates the potential of various machine learning algorithms as tools for air quality (AQ) episode forecasting and assesses them – given the corresponding domain-specific requirements – using an evaluation procedure, tailored to the task at hand. Among the algorithms employed in the experimental phase, our main focus is on ZCS-DM, an evolutionary rule-induction algorithm specifically designed to tackle this class of problems – that is classification problems with skewed class distributions, where cost-sensitive model building is required.

Overall, we consider this investigation successful, in terms of its aforementioned goals and constraints: obtained experimental results reveal the potential of rule-based algorithms for urban AQ forecasting, and point towards ZCS-DM as the most suitable algorithm for the target domain, providing the best trade-off between model performance and understandability.

1. Introduction

Air quality (AQ) forecasting is among the most challenging real-world application domains for machine learning algorithms. The nature of the problem – with its inherent irregularities and non-linearity – along with the often encountered lack of sufficient or good quality data constitute a multi-faceted and computationally demanding problem. Moreover, the corresponding regulatory framework poses a number of important requirements unique to the domain of AQ, among which the need for assessment of air pollution levels in areas of interest plays a central role. It is important to note, though, that, according to AQ management legislation, "assessment" is defined not only in terms of foreseeing-forecasting the quality of the atmospheric environment in the future, but also as a means to complement information received via monitoring stations in the target areas.

In this context, the development of scientific knowledge in the field, accompanied by the developing requirements in environmental decision making in general, and urban AQ more specifically, present a considerable challenge for both decision makers and AQ modelers: addressing decision-making needs (in terms of incident definitions, forecasting and assessment) with the aid of available AQ models. In both Europe and the USA, this has led to the investigation and use of metrics for the assessment of AQ model performance, as a commonly accepted procedure to estimate the success in the operation of such models. Further, the accumulated AQ management and policy making experience, in parallel with scientific developments, have led to a reformulation of the regulatory framework, entailing a more detailed and multi-level definition of the incidents of interest. Exactly in this direction, and for the first time in Europe, an integration of legislative provisions concerning AQ assessment has been implemented, leading to the Clean Air for Europe Directive (CAFE Directive 2008/50).

The CAFE Directive mandates not only the provision of up-to-date information on ambient air pollutant concentrations, but also the identification of specific characteristics of exceedances. These characteristics include the type of the pollutant and the geographic

* Corresponding author. Tel.: +30 2310996349; fax: +30 2310996398.
E-mail addresses: fani@olympus.ee.auth.gr (F.A. Tzima), mitkas@eng.auth.gr (P.A. Mitkas), voukas@isag.meng.auth.gr (D. Voukantsis), kkara@eng.auth.gr (K. Karatzas).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.148

area of exceedances, along with the type of threshold that is expected to be exceeded. The latter requirement, thus, introduces an "event-episode" typology distinction from the model developers' point of view and formulates a difficult forecasting problem – aiming at the prediction of AQ episodes – where effective models need to be built from data with highly skewed class distributions and relatively sparse episodes. This situation, combined with the fact that episode classes are more important (i.e. entail greater misclassification costs) than non-episode ones, implicitly poses the need for domain-specific assessment processes, capable of mapping the intended fault-tolerance of prediction models onto the evaluation metrics to be used.

Regarding AQ forecasts, the CAFE Directive also states that information should be provided on the expected changes in pollution after the incident takes place (improvement, stabilization or deterioration), focusing on the reasons for those changes. This requirement actually points to the fact that, in parallel to the typology of forecasted events, procedures (AQ models) employed should also be able to provide explanations concerning the behavior of the incident. Knowledge needs, thus, to be extracted from each analyzed event, in order to support decision making.

Given the aforementioned requirements, it is evident that, due to their inherent characteristics, machine learning methods, and data mining methodologies in general, are particularly suitable for the domain of AQ: they can be used for both knowledge extraction and incident-oriented forecasting, with the ultimate goal being public information provision and, more importantly, decision-making support.

On this basis, our current study investigates the potential of various machine learning algorithms as tools for AQ episode forecasting and assesses them – given the corresponding domain-specific requirements – using an evaluation procedure, tailored to the task at hand. More specifically, in the frame of our study, a series of models are developed for forecasting PM_{10} and O_3 concentration levels in the Metropolitan area of Thessaloniki, Greece. In the model development phase, 5 algorithms are employed: (i) Multi-Layer Perceptron (MLP) and (ii) Support Vector Machines (SVM-SMO), following the neural network based approach traditionally used in AQ forecasting, (iii) the well known tree-induction algorithm C4.5, (iv) the evolutionary rule induction algorithm ZCS-DM and (v) Logistic Regression that is included as a baseline for algorithm comparison. Among the algorithms used in this study our main focus is on ZCS-DM (Tzima & Mitkas, 2008), an evolutionary rule-induction algorithm that has specifically been designed to tackle classification problems with skewed class distributions, where cost-sensitive model building is required.

Following a modest parameter-tuning phase, the algorithms are evaluated based on two versions of the kappa statistic, the traditional one and a weighted version taking into account the various misclassification costs as they are perceived by decision makers. The statistical significance of the obtained results is also assessed using appropriate methods and, thus, providing not only a sound basis for algorithm comparisons, but also evidence for communicating the models performance and trustworthiness to decision makers. This communication is crucial to achieving the transition from theory to actual deployment and use of the developed models in the www.airthess.gr system, an environmental early warning system already in place in Thessaloniki.

The remainder of this paper is structured as follows: Section 2 describes the process of identifying the requirements of the target domain, namely urban AQ forecasting, and translating them into concrete modeling decisions for this study's experimental phase. Section 3 introduces the study area and presents the procedure used to formulate the corresponding pollutant time-series into datasets appropriate for the tasks at hand. The last part of Section 3 provides a brief introduction to the algorithms employed in this study, namely ZCS-DM, MLP, SVM, Logistic Regression and C4.5. Section 4 presents and discusses the obtained experimental results, while Section 5 concludes this work with some insights on the overall performance of the studied algorithms, as well as the degree of fulfillment of our investigation's initial objectives. Future research directions are also outlined, along with a discussion of our methodology's applicability to problem settings that share the characteristics/requirements identified in this study for the domain of AQ.

2. From domain-specific requirements to specific modeling decisions

According to EU legislation, the objective of AQ assessment models is, or at least should be, threefold: (i) providing the public with up-to-date information on ambient air pollutant concentrations, (ii) providing information of explanatory nature, regarding episodes after their occurrence, and (iii) forecasting episodes of poor AQ, thus assisting decision-makers in the difficult task of regulating urban AQ. It is evident that the latter two tasks are far from trivial, as they require an in-depth understanding of the interrelationships, dependencies and behavior of parameters of interest, as well as the integration of multiple pollutants having different spatio-temporal behaviors, physiochemical characteristics and limit values. Given the complexity of these processes, it is of paramount importance to relieve decision-makers of the burden of processing all available data and information and provide them with specific "summarized" and valuable knowledge. This knowledge has to be:

1. *easily interpretable by a human expert*, i.e., in rule-based or tractable mathematical form,
2. *trusted*, i.e., subjected to a thorough evaluation procedure and, possibly, assessed against existing knowledge,
3. *actionable*, i.e., easily translated to services and/or actions in actual decision making situations, and
4. *deployable*, i.e., integratable with existing AQ management systems, such as early warning systems providing information services to the public.

In light of the above requirements, a number of important modeling decisions has to be made when developing an AQ forecasting system, ranging from algorithm selection to evaluation methodologies and model interpretation processes employed.

2.1. Algorithm selection

The *interpretability* requirement for developed models has been largely neglected in previous work aiming at the prediction of urban AQ: neural networks (e.g. Corani, 2005; Hooyberghs, Mensink, Dumont, Fierens, & Brasseur, 2005; Nunnari, 2006) and classical regression techniques (e.g. Chaloulakou, Assimacopoulos, & Lekkas, 1999; Kaprara, Karatzas, & Moussiopoulos, 2001) have been widely used for this task, providing "black-box" prediction models that lack the desired level of understandability. Rule-based or tree-structured prediction models, on the other hand, have been less popular, although they usually entail a tractable number of rules, possibly hierarchically ordered, and following the traditional production system form (IF *conditions* THEN *prediction*).

Aiming at filling this gap, our current work investigates the potential of tree- and rule-based classifiers in the domain of urban AQ forecasting. Traditional algorithms, namely Neural Networks (NNs) and regression techniques, are also included in the model development phase, thus providing the basis for a comparison of the "intuitive" rule-based models with the "best-practice" standards of the domain.[1]

2.2. Evaluation methodology

The choice of the evaluation methodology to be employed is crucial not only in terms of providing concrete evidence on algorithm comparisons, but also to establish *decision-makers trust* in the developed models.

Regarding the aspect of assessing model 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 Prechelt's assessment (Prechelt, 1996) 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 urban AQ 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 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.

More specifically, in order to evaluate the statistical significance of the measured differences in algorithm performance, we have used the procedure suggested by Demšar (2006) for robustly comparing classifiers across multiple datasets. This procedure involves the use of the Friedman test (Friedman, 1937) to establish the significance of the differences between classifier ranks and, potentially, a post hoc test to compare classifiers to each other. In our case, the primary goal was to compare the performance of *all* employed algorithms to each other, thus, the Nemenyi test (Nemenyi, 1963) was selected as the appropriate post hoc test.

2.3. Evaluation metrics

The use of domain-specific performance indexes is of paramount importance not only for establishing the decision-makers trust in the developed models, but also for transforming extracted knowledge into *services and actions deployable in actual decision-making situations and/or systems*. This need is also mirrored in the "event-episode" typology introduced by recent AQ legislation, which implicitly calls for carefully designed evaluation metrics, mapping the intended fault-tolerance of predictions and, thus, able to distinguish the cases of interest in actual AQ management scenarios.

In our case, the final assessment of algorithm (and prediction model) performance is based on Cohen's Kappa (Cohen & April, 1960), defined as:

$$K = \frac{P_0 - P_C}{1 - P_C}, \tag{1}$$

where P_0 is the total agreement probability, and P_C is the agreement probability which is due to chance. Although initially introduced as a measure of agreement between observers of psychological behavior, it is evident how the statistic's applicability can be extended to the case of measuring the degree of agreement between a classification model's predictions and reality.

Comparing the performance of classifiers using the kappa statistic is an arguably robust evaluation method, compensating for classifications that may be due to chance. However, this approach is less suitable for cost-sensitive applications, where some errors may be less acceptable than others, or for decision problems with ordinal class values (Ben-David, 2008). This is due to the fact that, in both cases, considering all "misses" as of equal importance is non-realistic, thus resulting in performance evaluations that fail to quantify the actual error costs. The problem of predicting urban AQ as formulated in this study (i.e., as a classification problem) falls into both the aforementioned categories, as (i) for most operational forecasting cases, we are strongly interested in predicting exceedances, with "regular" values being less important and (ii) prediction class labels are ordered, making the "distance" of the misclassification at least as important as the error itself.

The above discussion makes the need for an error-weighting scheme evident: during the calculation of the kappa statistic, a mechanism should be employed to formulate penalties according to the intended fault-tolerance of the system. While weighting the errors in kappa can be done in many ways, one simple method is presented by Fleiss (1981) and defines that the value of the Weighted Kappa can be written as:

$$K_w = \frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_{i\bullet}P_{\bullet j}\right)}{\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_{i\bullet}P_{\bullet j}\right)}, \tag{2}$$

where w_{ij} denotes the weight of the count in the ith row and jth column of a confusion matrix, with $0 \leqslant w_{ij} \leqslant 1$, and $P_{i\bullet}$, $P_{\bullet j}$ are the marginal probabilities.

According to Eq. (2) the values of w_{ij} indicate the "strength" of agreement (or the cost of error), with $w_{ij} = 1$ expressing full agreement and $w_{ij} = 0$ full disagreement. Values across the main diagonal (that indicate full agreement) are usually equal to 1, while values outside the main diagonal ($w_{ij}, i \neq j$) decrease as the cost of disagreement increases.

Based on the above formulation of the weighted kappa coefficient, a plethora of weighting schemes can be employed, from linear to quadratic or even user-defined, when problem-dependent information about the cost of errors is available. The latter method was used in the current study, with a cost-matrix (Table 1) being defined according to our knowledge of the domain and mirroring the relative cost each of the model's misclassification has. More specifically:

- False "High alarms" entail a cost of unnecessary measures equal to 10
- False "Very High alarms" entail a cost of measures equal to 20
- Missed "High alarms" entail a cost (e.g. for public health) equal to 30

Table 1
Cost matrix used in the calculation of the weighted kappa coefficient. For the case of O_3 maximum daily values, only the top-left 3 × 3 cost values are taken into account.

Actual/predicted	Low	Medium	High	Very high
Low	0	5	10 + 10	15 + 20
Medium	5	0	5 + 10	10 + 20
High	10 + 30	5 + 30	0	5 + 20 + 30
Very High	15 + 90	10 + 90	5 + 10 + 90	0

[1] For more details on the specific algorithms used in the model development phase, see Section 3.3.

- Missed “Very High alarms” entail a cost equal to 90
- Misclassification costs, in general, are proportional to the distance between the actual and predicted class values, with every “step” away from the actual class entailing a cost equal to 5.

It is important to note that given the user-defined cost matrix (Table 1), the actual values of w_{ij} to be used in Eq. (2) are calculated by:

$$w_{ij} = \frac{max(cost) - cost_{ij}}{max(cost)}. \tag{3}$$

3. Materials and methods

3.1. Study area and target pollutants

Thessaloniki is the second largest city of Greece (more than one million inhabitants) and one of the most densely populated cities in Europe, accounting for approximately 17,800 inhabitants per km^2 (population statistics of 2001 for the municipality of Thessaloniki). Its complex coastal formation, in combination to the near-by mountainous areas, forms a very complex land use and orography pattern that favors local circulation systems. Thus, the formation and transport of pollutants are heavily influenced by the local meteorological and topographic characteristics (Güsten et al., 1997), which is the case in many coastal urban areas around the world.

AQ modeling with the aid of Computational Intelligence (CI) methods has been sparsely applied in Thessaloniki, aiming at analyzing air quality data and constructing forecasting models. Slini, Kaprara, Karatzas, and Moussiopoulos (2006) made use of linear regression, classification and regression trees and Artificial Neural Networks (ANNs) in an effort to forecast PM_{10} concentrations for Thessaloniki. Furthermore, Karatzas and Kaltsatos (2007) applied ANNs for the prediction of O_3 and NO_3 in Thessaloniki, while Tzima, Karatzas, Mitkas, and Karathanasis (2007) applied an arsenal of CI methods to investigate and forecast hourly PM_{10} concentration values in the same area.

The experimental setting proposed in our current study, concerns the operational prediction of mean PM_{10} and maximum O_3 concentrations in the Metropolitan Area of Thessaloniki. These pollutants (mainly PM_{10} and less O_3) are the main reasons of air quality problems in the area (Moussiopoulos, Papalexiou, & Sahm, 2006; Slini et al., 2006), yet they are of very different nature and origin:

- *Particulate matter* refers to a category of pollutants, further classified on the basis of their mean aerodynamic diameter and of the state that the are in. One of the “traditional” ones is PM_{10}, i.e. particulate matter of solid state and of mean diameter in the order of 10 μ m. This is a pollutant that is directly emitted by combustion processes and by traffic, while in some regions it is also produced as the result of mechanical degradation of the road surface and of winter tires. The criterion applied for assessment in the European Union is the daily mean concentration, and the limit value used equals 50 μ g/m^3, not to be exceeded more than 35 times per calendar year. Another criterion exists, concerning the mean annual value, which is 40 μ g/m^3 not to be exceeded.
- *Ozone (O_3)* is a pollutant not directly emitted, but produced in the atmosphere as the result of the change in the chemical balance of the atmospheric air, due to the existence of other pollutants. Ozone is a pollutant that has a very strong photochemical profile, and in addition can “travel” with the aid of atmospheric air. The criterion applied for assessment is the highest 8 h mean of hourly values, calculated as a running average; a set of 24 values should be calculated per day, each representing the 8 h average of time intervals ending from 01:00 to 24:00 of the day of reference. The target value is 120 μ g/m^3 not to be exceeded more than 20 days per calendar year. It is worth noting that the World Health Organization (WHO) has just introduced a new guideline value (equal to 100 μ g/m^3), a procedure typical in the domain of AQ management, resulting from updated scientific evidence concerning consequences of polluted air to man and the ecosystem.

3.2. Air quality time-series

Available atmospheric data in the Thessaloniki metropolitan area include hourly measurements of air quality and meteorological parameters, recorded by a network of monitoring stations operated by the Directorate of Environment and Land Planning of the Region of Central Macedonia, Greece (Fig. 1). In the present paper, we have used hourly atmospheric time series data, combined with seasonal information for estimating PM_{10} mean (in 3 stations) and O_3 maximum (in 7 stations) daily concentration levels. The data covered a time period of 8 years (2000–2008). Table 2 presents the available AQ and meteorological parameters for each one of the 7 monitoring stations taken under consideration.

In the data pre-processing phase, daily statistics of all prediction variables were calculated, i.e., minimum, maximum and mean values, while only the mean daily value was calculated for wind direction. Meteorological data were also pre-processed, so that they are taken from one day ahead, corresponding to the operational weather forecast. Additionally a series of seasonal variables (season, month, week, day of year, weekday, weekend) were also considered as input for the forecasting models. A modest feature selection phase, based on Information Gain (Cover & Thomas, 1991) was applied in order to prune the variable space, resulting in two distinct sets of input variables (Table 3), corresponding to the forecasting of PM_{10} mean (18 attributes) and O_3 maximum daily values (16 attributes).

The prediction models built are intended for use as operational tools for regulating AQ – identifying possible exceedances, issuing of alarms and measures etc. Thus, numerical values for the class variable were transformed into nominal ones, in accordance with

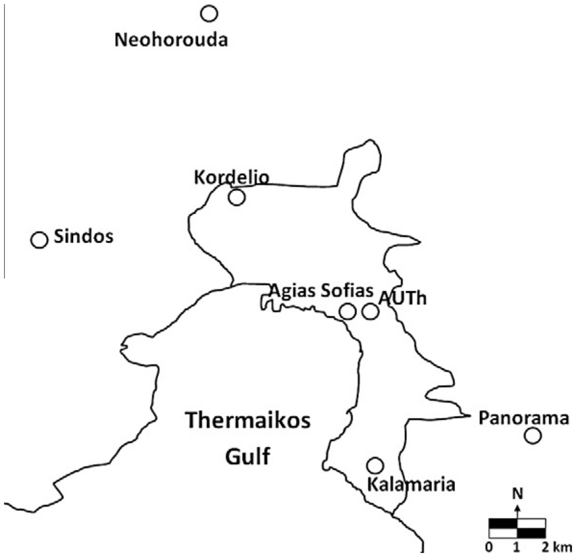


Fig. 1. AQ monitoring stations in the metropolitan area of Thessaloniki. The network is operated by the Directorate of Environment and Land Planning of the Region of Central Macedonia.

Table 2
Available time-series data per monitoring station. [Temp: Temperature, RH: Relative Humidity, WS: Wind Speed, WD: Wind Direction].

Station	AQ parameters	Meteorological parameters
Agias Sofias	CO, NO_2, O_3, PM_{10}, SO_2	Temp, RH, WS, WD
AUTh	NO_2, O_3, SO_2	Temp, RH, WS, WD
Kalamaria	CO, NO_2, O_3, SO_2	Temp, RH, WS, WD
Kordelio	CO, NO_2, O_3, PM_{10}, SO_2	Temp, RH, WS, WD
Neohorouda	NO_2, O_3	Temp, RH, WS, WD
Panorama	NO_2, O_3, PM_{10}	Temp, RH, WS, WD
Sindos	CO, NO_2, O_3, PM_{10}, SO_2	Temp, RH, WS, WD

Table 3
Selected sets of the input variables for PM_{10} mean and O_3 maximum daily values forecasting. [Temp: Temperature, RH: Relative Humidity, WS: Wind Speed, WD: Wind Direction].

PM_{10} mean daily values	O_3 maximum daily values
Month	Season
DayOfYear	Month
CO (mean, max)	Week
NO_2 (mean, max)	Day of year
O_3 (min)	O_3 (mean, max)
PM_{10} (mean, min, max)	Temp (mean, min, max)
WS (mean, min, max)	RH (mean, min)
WD	Temp Forecast (mean, min, max)
WS Forecast (mean, min, max)	RH Forecast (mean, min)
WD Forecast	

Table 4
Scales used for converting PM_{10} mean and O_3 maximum daily concentration values to nominal values.

Pollutant\ Scale	Low	Medium	High	Very high
PM_{10}	[1,50)	[50,90)	[90,110)	⩾110
O_3	[1,180)	[180,240)	⩾ 240	–

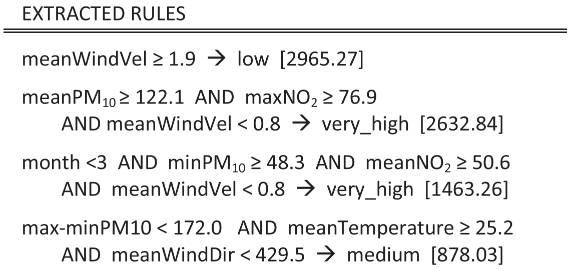
relevant technical guidelines defining the scales shown in Table 4. It is also worth noting that the scales used are in line with the operational alarm issuing policy currently employed by the Prefecture of Thessaloniki, that is responsible for regulating AQ in the area.

3.3. Methods

A series of models were developed using 5 algorithms: MLP and SVM-SMO, following the neural network based approach traditionally used in AQ forecasting, the well known tree-induction algorithm C4.5, the evolutionary rule-induction algorithm ZCS-DM and Logistic Regression that was included as a baseline for algorithm comparison. For all algorithms the implementation used was the one provided by the machine learning tool WEKA (Witten & Frank, 2005), with the exception of ZCS-DM for which we used our custom "in-house" implementation.

3.3.1. ZCS-DM

ZCS-DM belongs to a class of machine learning methods, named Learning Classifier Systems (LCS), that were originally proposed by Holland (1975) and are capable of tackling both single-step and sequential decision problems (Sigaud & Wilson, 2007). Although such algorithms entail a number of quite elaborate mechanisms, in the following we will only provide a high level description of ZCS-DM, focusing on the components necessary for our current investigation of a single-step decision task, namely the prediction of daily pollutant concentrations in the Metropolitan Area of Thessaloniki. For a more in-depth description of the algorithm, as well as recommendations for parameter settings the reader is referred to Tzima and Mitkas (2008) and Wilson (1994).


EXTRACTED RULES

meanWindVel ≥ 1.9 → low [2965.27]

$meanPM_{10}$ ≥ 122.1 AND $maxNO_2$ ≥ 76.9
AND meanWindVel < 0.8 → very_high [2632.84]

month <3 AND $minPM_{10}$ ≥ 48.3 AND $meanNO_2$ ≥ 50.6
AND meanWindVel < 0.8 → very_high [1463.26]

max-minPM10 < 172.0 AND meanTemperature ≥ 25.2
AND meanWindDir < 429.5 → medium [878.03]

Fig. 2. Example of rules extracted by ZCS-DM.

ZCS-DM employs a population $R = \{R_1, R_2, \ldots, R_N\}$ of gradually evolving, cooperative rules, each encoding a fraction of the problem domain and, thus, collectively forming the overall solution to the target problem (the final rule set to be used for AQ forecasting in our case). Rules in ZCS-DM are represented in the production-system form of "IF *conditions* THEN *prediction*" (Fig. 2) and are encoded over the ternary alphabet {0, 1, #}, in line with the encoding traditionally used in Genetic Algorithms (GAs). The symbol # (usually termed as a "wildcard" or a "don't care") allows for generalization in the rule condition part, such that both inputs 11 and 10 are matched by the rule condition 1#. No generalization occurs in the decision part – decisions are discrete and usually integer-valued. Associated with each rule, there is a scalar strength value expressing its expected reward i.e., its usefulness for the prediction task.

Through each training cycle, ZCS-DM receives a binary encoded data instance (a vector of prediction variable values), determines an appropriate response based on a rule (or a set of rules) whose condition matches the input, and produces a classification decision. Successful classification of an instance is associated with a scalar reward R apportioned to the system rules according to a reinforcement scheme. Thus, at each discrete time-step, the system follows a cycle of *performance*, *reinforcement* and *discovery* component activation, with the last component employing two rule discovery mechanisms: (i) a steady-state GA and (ii) a covering operator that is activated when there is no (competent) rule matching the current input.

3.3.2. Multi-layer perceptron (MLP)

Multi Layer Perceptrons (MLPs), consisting of simple processing units (neurons) massively interconnected to each other, have been successively applied for modeling complex processes in various application domains, including the domain of AQ forecasting (Gardner & Dorling, 1998; Kukkonen et al., 2003). They are suitable for both, regression and classification tasks. In the latter case, the number of neurons at the output layer is set to be equal to the classes of the target variable, while the forecasted class is indicated by a binary output (0: not predicted, 1: predicted). An extensive discussion of MLPs, and Artificial Neural Networks (ANNs) in general, may be found in Haykin (1998).

The MLP models developed in our current work consist of one hidden layer, whose number of neurons was determined experimentally: for the case of PM_{10} mean daily values it was set to 12, while for the O_3 maximum daily values to 8. Furthermore, the maximum number of epochs was set to 300 and combined with an early stopping criterion (validation set proportion of 10% of the training set), in order to avoid overfitting.

3.3.3. Support vector machines-SMO (SVM-SMO)

Support Vector Machine (SVM), introduced by Vapnik (1999), is a machine learning method capable of performing classification, regression and function approximation tasks. There are several versions of the SVM algorithm and a series of parameters that need to be optimized, in order to achieve good generalization properties and performance. In this paper, the Sequential Minimal Optimization (SMO) implementation provided by WEKA was used and two parameters were optimized: (i) the capacity constant C and (ii) the variance of the Gaussian kernel parameter γ. The capacity constant, specifying the penalization of error, was set to be equal to 10 for the case of PM_{10} forecasting and 60 for the case of O_3. Furthermore, the following radial basis kernel function was used to map the input space:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \tag{4}$$

where γ denotes the variance of the Gaussian kernel. For both target variables, the optimal value of γ was experimentally found to be equal to 1.

3.3.4. C4.5

C4.5 is an improvement of the ID3 algorithm, introduced by Quinlan (1993), that builds decision trees using the concept of Information Gain. C4.5 can be used only for classification tasks and involves several parameters that can be tuned to achieve optimal performance. An important parameter is the confidence factor, which affects pruning: smaller confidence factor values result in increased pruning and, thus, smaller trees.

The J48 implementation of C4.5 (provided by WEKA) was used in this study, while the confidence factor was chosen to be equal to 0.025 for PM_{10} forecasting and 0.02 for O_3 after an optimization process.

3.3.5. Logistic regression

Finally, the well known Logistic Regression algorithm (and more specifically the Multinomial Logistic Regression implementation provided by WEKA) was used as a baseline for algorithm comparison.

4. Experiments and results

The experimental procedure employed in our current investigation aimed at evaluating the studied algorithms' performance relative to the each other, following a modest tuning phase for selecting the optimum parameter set per algorithm. A secondary objective was the investigation of rule-based algorithms' – namely C4.5 and ZSC-DM – potential for the AQ forecasting task, since this category of models better fulfills the requirement for model understandability identified during the requirements gathering phase. The procedure followed to reach conclusions regarding the final model to be selected for operational forecasting would also have to use domain-specific evaluation metrics and a robust evaluation procedure, providing concrete evidence for the communication of results to the interested parties.

Thus, based on the modeling choices and procedures outlined in Section 2, the 10-fold cross validation method was used for all experiments, while the final assessment of algorithm performance was based on two versions of the kappa statistic: (i) the traditional one and (ii) a weighted version, taking into account the various misclassification costs, as perceived by decision makers.

The numerical results obtained for the kappa and the weighted kappa coefficients are given in Tables 5a and 6a, respectively. In both tables, the best kappa value (per dataset) is presented in bold, while average ranks are calculated both per pollutant (PM_{10} and O_3 Av. Rank) and over all the datasets (Overall rank).

The highest performances, in terms of the conventional kappa index, were obtained with SVM and ZCS-DM models, with the two algorithms having an overall average rank of 1.75 and 2.5, respectively. SVM and ZCS-DM also achieved the best average rank for the PM_{10} and O_3 forecasting tasks, respectively.

When assessing performance in terms of the weighted kappa index, Logistic Regression clearly outperforms all its rivals in the PM_{10} forecasting cases, with an average rank of 1.50. On the other hand, ZCS-DM performs better than Logistic Regression (O_3 Av. Rank: 2.857) and SVM (O_3 Av. Rank: 2.143) in the O_3 forecasting cases, with the highest average rank of 2.00. The same ranking holds for the overall forecasting task, with ZCS-DM outperforming its rivals and producing models that better map the intended fault-tolerance of the prediction task. We consider these findings indicative of ZCS-DM's ability to forecast rare and extreme AQ episodes, unlike statistical methods that typically produce models "averaging" over the observations, thus being less suitable for the prediction of extreme values of phenomena.

In general, we consider this initial investigation successful, as the results obtained reveal the potential of interpretable

Table 5
Model comparison based on the kappa coefficient.

(a) Kappa values per dataset and corresponding algorithm ranks (in parentheses)

Algorithms Datasets	C4.5	Logistic Regression	MLP	SVM	ZCS-DM
PM_{10}–*Agias Sofias*	0.432 (4)	0.443 (3)	0.445 (2)	**0.468** (1)	0.417 (5)
PM_{10}–*Kordelio*	0.434 (3.5)	0.470 (2)	0.434 (3.5)	**0.486** (1)	0.419 (5)
PM_{10}–*Sindos*	0.458 (3)	0.450 (5)	0.457 (4)	**0.468** (1)	0.464 (2)
PM_{10} Av. Rank	3.500	3.333	3.167	**1.000**	4.000
O_3–*Agias Sofias*	0.343 (5)	0.425 (4)	0.483 (2)	0.427 (3)	**0.492** (1)
O_3–*AUTh*	0.697 (5)	0.707 (3)	0.703 (4)	0.708 (2)	**0.721** (1)
O_3–*Kalamaria*	0.490 (5)	0.553 (4)	0.602 (2.5)	0.602 (2.5)	**0.607** (1)
O_3–*Kordelio*	0.731 (3)	0.728 (4)	0.710 (5)	0.737 (2)	**0.739** (1)
O_3–*Neohorouda*	0.734 (5)	**0.751** (1)	0.736 (4)	0.745 (3)	0.749 (2)
O_3–*Panorama*	0.702 (4)	0.722 (2)	0.721 (3)	**0.727** (1)	0.678 (5)
O_3–*Sindos*	0.635 (5)	0.671 (3)	0.660 (4)	**0.684** (1)	0.675 (2)
O_3 Av. Rank	4.571	3.000	3.500	2.071	**1.857**
Overall Rank	4.250	3.100	3.400	**1.750**	2.500

(b) Differences in overall algorithm ranks (based on kappa) for the Nemenyi post hoc test

	Overall rank				O_3 Av. Rank			
	C4.5	Log. R.	MLP	SVM	C4.5	Log. R.	MLP	SVM
Log. R.	1.15				1.57			
MLP	0.85	−0.30			1.07	−0.50		
SVM	2.50	1.35	1.65		2.50	0.93	1.43	
ZCS-DM	1.75	0.60	0.90	−0.75	2.71	1.14	1.64	0.21

Table 6
Model comparison based on the weighted kappa coefficient.

(a) Weighted-kappa values per dataset and corresponding algorithm ranks (in parentheses)

Algorithms Datasets	C4.5	Logistic Regression	MLP	SVM	ZCS-DM
PM_{10}–*Agias Sofias*	0.299 (5)	0.347 (2)	0.335 (4)	0.346 (3)	**0.365** (1)
PM_{10}–*Kordelio*	0.445 (5)	**0.477** (1.5)	0.452 (4)	0.469 (3)	**0.477** (1.5)
PM_{10}–*Sindos*	0.337 (2)	**0.339** (1)	0.326 (3)	0.317 (4)	0.238 (5)
PM_{10} Av. Rank	4.000	**1.500**	3.667	3.333	2.500
O_3–*Agias Sofias*	0.320 (5)	0.390 (4)	0.455 (2)	0.400 (3)	**0.468** (1)
O_3–*AUTh*	0.647 (5)	**0.682** (1)	0.651 (4)	0.659 (3)	0.672 (2)
O_3–*Kalamaria*	0.424 (5)	0.472 (4)	0.531 (3)	0.542 (2)	**0.548** (1)
O_3–*Kordelio*	0.611 (3)	0.603 (4)	0.591 (5)	**0.640** (1)	0.620 (2)
O_3–*Neohorouda*	0.619 (4)	0.679 (2)	0.632 (3)	**0.680** (1)	0.615 (5)
O_3–*Panorama*	0.603 (5)	0.634 (2)	0.633 (3)	0.624 (4)	**0.647** (1)
O_3–*Sindos*	0.629 (5)	0.666 (3)	0.655 (4)	**0.678** (1)	0.669 (2)
O_3 Av. Rank	4.571	2.857	3.429	2.143	**2.000**
Overall Rank	4.400	2.450	3.500	2.500	**2.150**

(b) Differences in overall algorithm ranks (based on weighted kappa) for the Nemenyi post hoc test

	Overall Rank				O_3 Av. Rank			
	C4.5	Log. R.	MLP	SVM	C4.5	Log. R.	MLP	SVM
Log. R.	1.95				1.71			
MLP	0.90	−1.05			1.14	−0.57		
SVM	1.90	−0.05	1.00		2.43	0.71	1.29	
ZCS-DM	2.25	0.30	1.35	0.35	2.57	0.86	1.43	0.14

rule-based models – developed using the ZCS-DM algorithm – in complex prediction (classification) tasks, such as urban AQ forecasting. Although the prediction accuracy of ZCS-DM was in some cases at moderate level (ZCS-DM kappa: 0.417–0.749), its performance was only slightly worse or equivalent to that of the traditional “black-box” MLP and Logistic Regression models and clearly better than the corresponding performance obtained with the standard C4.5 approach. Results are even more promising when cost-sensitive evaluation is applied, based on the weighted kappa statistic: ZCS-DM outperforms all rival algorithms in 5 out of the 10 datasets used in this study, with SVM and Logistic Regression being the only algorithms to achieve a better weighted kappa value, in 3 and 2 cases respectively.

4.1. Statistical evaluation of results

In order to evaluate the statistical significance of the measured differences in algorithm ranks, we have used the procedure suggested by Demšar (2006) for robustly comparing classifiers across multiple datasets. This procedure involves the use of the Friedman test, a non-parametric statistical test for evaluating the differences between more than two related sample means (Friedman, 1940) – where the related samples are, in our case, the performances of the $k = 5$ studied algorithms across the N target datasets. The null hypothesis being tested is that all classifiers perform the same and any observed differences are merely random. This is equivalent to testing whether the measured average ranks r_j ($j = [1,N]$) are significantly different from the expected mean rank $\bar{r} = \frac{k+1}{2} = 3$. The statistic used in the Friedman test is Iman and Davenport’s F_F (Iman & Davenport, 1980), that is distributed according to the F-distribution with $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom and defined as:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{5}$$

where

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j r_j^2 - \frac{k(k+1)^2}{4}\right]. \tag{6}$$

The procedure described above has been applied to the data in Tables 5a and 6a, which compare the five studied algorithms. As already stated, average ranks provide a fair comparison of the algorithms, revealing that, on average, ZCS-DM and SVM rank second based on both the kappa and the weighted kappa coefficients. Given the measured average ranks, the Friedman test checks whether the average ranks are significantly different form the mean rank $\bar{r} = 3$ expected under the null hypothesis. With five algorithms and 10 data sets (for the overall forecasting task), F_F is distributed according to the F-distribution with 4 and 36 degrees of freedom. The critical value of F (4,36) for $\alpha = 0.05$ is 2.63, so, with F_F values of 4.84 and 4.91 respectively, we reject the null hypothesis both when comparison is based on the conventional and the weighted kappa coefficients. When considering the two forecasting cases separately, the null hypothesis is only rejected for the subset of O_3 prediction cases ($N = 7, \alpha = 0.05$), but not for the PM_{10} ones ($N = 3$, various α values).

Having found that the measured average ranks are significantly different (at $\alpha = 0.05$), we can proceed with the post hoc tests that evaluate the relative performance of the studied algorithms against each other (i) for the O_3 and (ii) the overall forecasting task. The test selected in our case is the Nemenyi test (Nemenyi, 1963) that compares all classifiers to each other, and categorizes them in two “groups”, such that for classifiers of the same group, there cannot be any statistically significant assessment about their relative performance.

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

$$CD = q_\alpha\sqrt{k(k+1)/6N}, \tag{7}$$

where critical values q_α are those of the Studentized range statistic divided by $\sqrt{2}$ with a significance level of α and k degrees of freedom. The values of the critical difference for the two cases of applying the Nemenyi test (the O_3 and overall forecasting tasks) are given in Table 7 for significance levels $\alpha = 0.05$ and $\alpha = 0.1$.

Thus, based on the differences in algorithm ranks reported in Tables 5b and 6b, the application of the Nemenyi test reveals that:

Table 7
Critical Difference (CD) values for the Nemenyi post hoc test.

	$N = 7$	$N = 10$
$\alpha = 0.05$	1.522	2.133
$\alpha = 0.1$	1.372	1.922

- For the O_3 forecasting task ($\alpha = 0.05$), the performance of Logistic, SMO and ZCS-DM is significantly better than that of C4.5, when assessment is based both on the kappa and weighted kappa coefficients. Moreover, at $\alpha = 0.1$, the performance of SVM (based on kappa) and ZCS-DM (based on both kappa and weighted kappa) is significantly better than that of MLP.
- For the overall forecasting task and at $\alpha = 0.05$, SVM and ZCS-DM significantly outperform C4.5, when assessment is based on the kappa and the weighted kappa coefficients, respectively. Moreover, at $\alpha = 0.1$, the performance of Logistic is significantly better than that of C4.5, when assessing the corresponding models based on their kappa values.

From the above discussion, it is evident that the statistical evaluation of obtained results reinforces our conclusions about the potential of the studied algorithms in the target domain: in all studied forecasting tasks, ZCS-DM, our evolutionary rule-induction algorithm, performs comparably with traditional "black-box" NN- or regression-based models, and significantly better than the standard C4.5 approach. Moreover, in parallel to achieving its design objectives in the target cost-sensitive setting, ZCS-DM also takes into account the important requirement for models in a tractable rule-based form. By providing the best trade-off between model performance and understandability, it is thus a prime candidate for the choice of the most suitable algorithm for the task at hand.

5. Conclusions and future work

In this paper we have investigated the applicability and potential of several machine learning algorithms, in the task of predicting urban AQ. The objective of our investigation was two-fold and included (i) the identification of the requirements specific to the target domain and (ii) their transformation into concrete modeling choices for the experimental phase, aimed at developing robust and effective forecasting models, in line with the domain experts' view of the field. In this direction, and following a carefully designed experimental methodology, we pursued each of these goals, especially focusing on the aspect of robustly evaluating model performance in the target cost-sensitive setting.

More specifically, a series of models were developed for forecasting PM_{10} and O_3 concentration levels in the Metropolitan area of Thessaloniki, Greece, using 5 algorithms, namely MLP, SVM-SMO, C4.5, ZCS-DM and Logistic Regression. Following a modest parameter-tuning phase, the algorithms were evaluated based on two versions of the kappa statistic, the traditional one and a weighted version taking into account the various misclassification costs, as they are perceived by decision makers. The statistical significance of the obtained results was also assessed using appropriate methods and, thus, providing not only a sound basis for algorithm comparisons, but also evidence for communicating the models performance and trustworthiness to decision makers.

Overall, we consider our investigation successful, in terms of its aforementioned goals and constraints. Obtained results are encouraging, since all algorithms achieve acceptable performance on the prediction targets, being able to accurately forecast air pollution episodes and, thus, providing knowledge directly usable by environmental information systems and in accordance with the new legal and regulatory framework for AQ management in Europe. More importantly, rule-based models, that are in line with the decision makers requirement for interpretability, perform comparably with the NN- and regression-based models traditionally used in the domain of AQ. Thus, given the thorough evaluation procedure employed and taking into account the important requirement for models in a tractable rule-based form, we consider that our choice of ZCS-DM as the most suitable algorithm for the task at hand provides the best trade-off between model performance and understandability.

Of course, several issues remain open for future investigations that should be focused on (i) proving the actionability of the developed models in practice, through their deployment in actual decision-making situations in the www.airthess.gr early warning system, and (ii) devising procedures that will "translate" extracted knowledge into usable information about the mechanisms governing forecasted episodes and their spatio-temporal evolution.

Another important issue, on the other hand, is the investigation of our methodology's applicability in domains other than that of AQ forecasting. Given the characteristics of the studied domain and the structured approach we have applied in this case, we believe that our methodology can be effectively applied to other domains, sharing the characteristics of AQ forecasting, either concerning *the nature of the problem* – skewed class distributions, focus on extreme cases/ episode forecasting, need for interpretable models, or *the evaluation process employed* – cost-sensitive evaluation taking into account the end-user's view of the field, thorough assessment of developed models providing concrete evidence on their performance and trustworthiness.

Acknowledgments

The first 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 cofinanced 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). The authors also acknowledge the Directorate of Environment and Land Planning of the Region of Central Macedonia, Greece for providing with the atmospheric data used in this paper.

References

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications, 34*(2), 825–832.

Chaloulakou, A., Assimacopoulos, D., & Lekkas, T. (1999). Forecasting daily maximum ozone concentrations in the Athens basin. *Environmental Monitoring and Assessment, 56*(1), 97–112.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Corani, G. (2005). Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modeling, 185*, 513–529.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley and Sons, Inc.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. NY, USA: Wiley.

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, 32*(200), 675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of *m* rankings. *The Annals of Mathematical Statistics, 11*(1), 86–92.

Gardner, M., & Dorling, S. (1998). Artificial neural network (the multilayer perceptron) – A review of applications in the atmospheric sciences. *Atmospheric Environment, 6*, 2627–2636.

Güsten, H., Heinrich, G., Mönnich, E., Weppner, J., Cvitas, T., Klainsinc, L., et al. (1997). Thessaloniki '91 Field measurement campaign-II. Ozone formation in the greater Thessaloniki area. *Atmospheric Environment, 31*(8), 1115–1126.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Cambridge, MA, USA: MIT Press.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., & Brasseur, O. (2005). A neural network forecast for daily average PM_{10} concentrations in Belgium. *Atmospheric Environment, 39*(18), 3279–3289.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the Friedman statistic. *Communications in Statistics, A9*(6), 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. In *Proceedings of the VII international conference on harmonization within atmospheric dispersion modelling for regulatory purposes* (pp. 193–196).

Karatzas, K. D., & Kaltsatos, S. (2007). Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. *Simulation Modelling Practice and Theory, 15*(10), 1310–1319.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., et al. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment, 37*(32), 4539–4550.

Moussiopoulos, N., Papalexiou, S., & Sahm, P. (2006). Wind flow and photochemical air pollution in Thessaloniki, Greece. Part I: Simulations with the European zooming model. *Environmental Modelling and Software, 21*(12), 1741–1751.

Nemenyi, B. (1963). Distribution-Free Multiple Comparisons. Ph.D. Thesis. Princeton University.

Nunnari, G. (2006). An improved back propagation algorithm to predict episodes of poor air quality. *Soft Computing, 10*(2), 132–139.

Prechelt, L. (1996). A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. *Neural Networks, 9*(3), 457–462.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*, 317–327.

Sigaud, O., & Wilson, S. W. (2007). Learning classifier systems: A survey. *Soft Computing, 11*(11), 1065–1078.

Slini, T., Kaprara, A., Karatzas, K., & Moussiopoulos, N. (2006). PM_{10} forecasting for Thessaloniki, Greece. *Environmental Modelling and Software, 21*(4), 559–565.

Tzima, F. A., Mitkas, P. A., 2008. ZCS Revisited: A zeroth-level classifier system for data mining. In *Proceedings of the 2008 IEE international conference on data mining workshops* (pp. 700–709).

Tzima, F. A., Karatzas, K. D., Mitkas, P. A., Karathanasis, S. (2007). Using data-mining techniques for PM_{10} forecasting in the metropolitan area of Thessaloniki Greece. In *IJCNN. IEEE* (pp. 2752–2757).

Vapnik, V. N. (1999). *The nature of statistical learning theory (information science and statistics)*. Springer.

Wilson, S. W. (1994). ZCS: A zeroth-level classifier system. *Evolutionary Computation, 2*(1), 1–18.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.