07JEI00091
1726-2135/1684-8799

www.iseis.org/jei

Journal of Environmental Informatics 9(2) 100-107 (2007)

Knowledge Discovery for Operational Decision Support in Air Quality Management

I. N. Athanasiadis[1*] and P. A. Mitkas[2]

[1]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, Switzerland
[2]Aristotle University of Thessaloniki, Thessaloniki, Greece

ABSTRACT. Operational decision-making in air quality management systems requires intense efforts for assessing monitored data streams on time. In contrary with the previous works, which focus on air quality forecasting, this paper concentrates on near real time air quality assessment. Data uncertainty problems associated with environmental monitoring networks bring forth issues such as measurement validation and estimation of missing or erroneous values, which are critical for taking trustworthy decisions in a timely fashion. A remedy to these problems is proposed through knowledge discovery techniques. By employing classification techniques, an empirical approach is presented for supporting the decision making process involved in an environmental management system that monitors ambient air quality and triggers alerts when incidents occur. Specifically, exhaustive experiments with large, real world data-sets have resulted to trustworthy predictive models, capable operational decision-making for measurement validation and estimation of missing or erroneous data. The outstanding performance of the induced predictive models signifies the added value of using data-driven approaches in operational air quality assessment.

Keywords: Air quality management, erroneous data estimation, knowledge discovery, measurement validation, operational decision support

1. Introduction

Decision-making in environmental operational centers is a demanding activity that engages greater than ever efforts, as the public demand for information is increasing and the monitoring networks are expanding. In this work, we focus our attention on the decision-making processes involved in air quality operational centers. Specifically, the main research goal of this paper is to demonstrate how data-driven approaches can be utilized for operational decision-making related to urban air quality. Even if the actual processes occurring in the atmosphere are extremely complex, the data driven approaches seem to come out with trustworthy models that can be used for operational decision-making.

Nowadays, environmental monitoring networks, including sensor recording, and radar and satellite images have been established worldwide in order to observe the conditions of the natural environment. As hardware and communication costs are decreasing, environmental monitoring networks are expanding at exponential rates, generating, faster than ever before, vast volumes of raw data, which need to be processed, analyzed, comprehended, and stored. Typically, Environmental Management Information Systems (EMIS) are occupied with gathering, preprocessing and integrating data-streams recorded by monitoring networks. A common EMIS installation involves the fusion into a central database of all data sensed at distributed locations. Until recently, all recorded data were meant for environmental scientists, who are engaged in off-line studies and post-processing activities in their effort to better understand the natural phenomena involved and forecast potentially harmful incidents. However, during the last period there has been a transition in environmental monitoring systems: Growing public interest in environmental protection and sustainable development has emerged the need for the diffusion of environmental information to all social parties. It is evident that public awareness affects the response of the involved stakeholders and the effectiveness of prevention measures. Thus, legislative acts in Europe and the US have deliberated environmental quality indicators, which need to be communicated to the public *on-time*, i.e. at the time incidents occur. As a consequence, *operational environmental assessment modules* (i.e. for near real time incident identification and reporting) have to be incorporated in EMIS.

Operational environmental assessment modules require an effective decision support at a near-real time frame, which practically means that the time required reaching a decision becomes a very important factor. Analytical air-quality models are far too complex for responding at an operational time window. Furthermore, the time constraint reveals two critical problems in operational decision-making: (a) the low quality or absence of sensed data, and (b) the changing conditions over long periods of time. In this context, quantitative data-driven decision support models are challenged by the difficulties in handling dynamic and uncertain features of real-world environmental systems. As Huang and Chang (2003) underlined, conditions for environmental management keep changing with time, demanding periodically updated decision support.

* Corresponding author: ioannis@athanasiadis.info

In this paper, we present how efficient decision support for operational environmental assessment and trustworthy information dissemination, can be realized by *learning from data*, using knowledge discovery techniques, for minimizing human intervention and taking decisions at the time incidents occur. The rest of the paper is structured as follows. Section 2 presents knowledge discovery terminology and provides with some background on prior experiences of applying knowledge discovery techniques in the air quality domain. Section 3 presents air quality assessment operational centers, air quality indicators and the decision-making processes involved. The fourth section introduces the knowledge discovery approach for operational air quality assessment and presents two predictive models for ozone measurement validation and ozone level estimation for erroneous and missing measurements. These models were cross-evaluated with data from three meteorological stations for a three year period, as analytically presented in Section five. Finally, results are discussed in the last section six were the conclusions of this study are drawn.

2. Knowledge Discovery and Previous Work

2.1. Terminology

Knowledge discovery is a broad term that has been used in computer science as an umbrella term that embraces a variety of techniques and procedures for inducing knowledge from data. The knowledge discovery process starts with specifying the decision models according to the problem at hand, and goes on with preparing and cleaning the original data, selecting suitable data mining algorithms, tuning and executing them for extracting knowledge, and, finally, interpreting the extracted knowledge patterns, report and reuse the results (Fayyad et al., 1996). In this respect, the term knowledge discovery describes the whole procedure for extracting knowledge patterns from data, as the focus is not limited on the application of some algorithm on a dataset (which is referred to as data mining). *Empirical approaches*, such as statistical regression, time-series analysis, artificial neural networks, case based reasoning, decision trees and rule induction algorithms are considered to be part of the data-mining algorithms available, and according to the task at hand, one or more appropriate algorithms can be considered for evaluation. The variety of data mining algorithms available for different decision tasks (as classification, clustering, regression, and time series analysis, etc.) is presented in reference book, as in Han and Kamber (2006), Klosgen et al. (2002), and Witten and Frank (1999).

2.2. Knowledge Discovery Applications on Air Quality Data

Earlier research work has dealt with EMS using knowledge discovery techniques mainly for forecasting purposes in the long or in the short term. Several models have been built for predicting incidents that may occur in the near future. For instance, the conventional statistical regression models (Bordignon et al., 2002; Huang and Smith, 1997; Kim and Guldmann, 2001) and time-series analysis (Chen et al., 1998) have been applied to predict ozone levels. Neural networks have been used for short-term ozone predictions (Ruiz-Suarez et al., 1995, Yi and Prybutoc, 1996), while case-based reasoning (Lekkas et al., 1994) and classification and regression trees (Kalapanidas and Avouris, 2001) have been employed for predicting air pollutant concentrations. In all aforementioned approaches, the decision making process related to incident forecasting has been successfully supported through the use of knowledge discovery techniques, such as statistical models, knowledge bases, case-based reasoning, classification trees, or artificial neural networks. A more detailed discussion on how knowledge discovery techniques have been applied on environmental data is provided in Gilbert et al. (2007).

In this work, the knowledge discovery techniques will be applied for supporting decision making processes involved in an EMIS, from a different perspective: Our main goal is *not to forecast* oncoming incidents, rather is *to assess on-time*, the monitored environmental conditions. This diversion in the point of view is a consequence of the emerging requirements of on line decision-making and rear real time reporting systems. This is the major contribution of this work that demonstrates the capabilities of empirical methods for operational decision-making in the ambient air quality domain. The development of data-driven decision strategies for successfully assessing ambient air quality at 'near real time' is presented through the exploitation of machine learning techniques.

3. Ambient Air Quality Assessment

3.1. Operational Air Quality Centers

Air quality depreciates in many cities, as a result of industrial activities and traffic emissions. For this reason, Air Quality Operational Centers have been established as monitoring networks in areas with potential air pollution problems. These networks sense atmospheric conditions and trace related measurements, such as the meteorological attributes and pollutant concentrations. Air Quality Operational Centers are responsible for processing all the recorded information and assess air quality. Certain indicators have been established in Europe and the US to determine air quality in urban areas, according to the European Directive on Ambient Air Quality (1996) and the US Clean Air Act (1990). Air pollutant concentration distinction in 'Air Quality Bands' has been applied to help the public associate pollution levels with possible health impacts. Air quality indicators issued by the European Commission are summarized in Table 1. In general, the calculation of air quality indicators is a simple, well-defined, straightforward procedure, as it involves the calculation of the average concentration in a certain time-frame. An in-depth discussion on air quality indicators and their association with human health can be found in Barratt (2000). It worths mentioning that EU Directives on Air Quality have delimited the 'Information' and 'Alert' levels for air pollutant concentrations, associated with these bands. Specifically, European Directive 92/72/EEC arranges to inform the public when warning and information threshold levels are exceeded.

3.2. Data Uncertainty in Air Quality Networks and Current Practice

In environmental monitoring networks, various 'sensor breakdown events', such as sensor malfunction, network delay, or noise, may lead to losses of or biased measurements. Consequently, all follow-up tasks including the identification of alerts are disabled or less credible, potential incidents cannot be identified at the time they occur, and human intervention is needed for substituting the missing measurements.

Data uncertainty inherited from monitoring networks affects the efficient calculation of the air quality indicators. The typical procedure followed by the majority of Air Quality Operational Centers involves human experts to overcome data uncertainties. Usually, environmental scientists are engaged to assess air quality at real time and to trigger alarms when 'Information' and 'Alert' levels are exceeded. The US Environmental Protection Agency suggests a data quality assurance procedure through sophisticated graphing systems that allow monitoring staff to quickly review data coming from the monitoring network (Hedges, 1999). To overcome such issues, flexible data analysis is supported through statistical tools in the London Air Quality Network (LAQN; Barratt, 2000), while meteorologists set criteria (expert rules) for validating data and making predictions in the Texas Natural Resource Conservation Commission (TCEQ).

It is evident that even if the calculation of air quality indicators is a simple, straightforward task, the preparatory activities involved for data validation and missing measurement estimation are complex processes, which typically undertaken by human experts. Even if recorded measurements are available through the network close to the time incidents occur, data validation and review is a struggling task that makes the whole procedures time-consuming. The European Environment Agency indicates that the validated data, reviewed by environmental scientists, are made available one to six months after measurement (Larssen and Hagen, 1996). In this context, it is valuable to automate the procedure with respect to time-pressure constraints.

3.3. The Overall Decision-Making Process for Operational Air Quality Assessment

As pointed out previously, two are the driving forces involved in EMIS for ambient air quality assessment. First comes the societal need for information, communicated at the time that an incident takes place. The second is data uncertainty, which is translated into an enormous workload for environmental scientists. In this context, an EMIS *operational decision-making module* is expected, not simply to calculate the air quality indicators, but to deal with data uncertainties and to adapt to an ever-changing environment.

The overall decision-making process in a generic automated EMS for ambient air quality assessment can be schematically represented as a fish diagram (Fig. 1). The starting point is to fuse all sensory inputs into the system. Then, a procedure of four decision-making steps follows. The first step is to validate incoming measurements. The second is to substitute invalid measurements, i.e. missing or erroneous ones. Finally comes, the calculation of formal alarms and the identification of custom alarms, for assessing air quality.

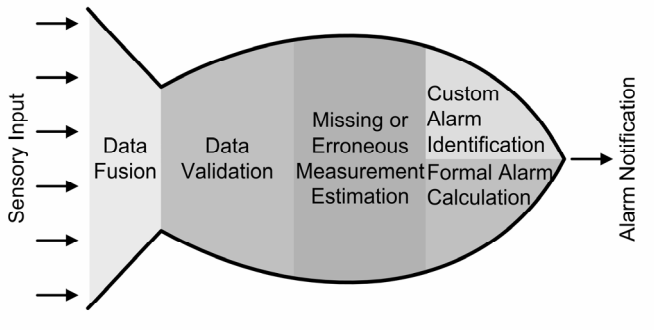

Figure 1. The decision making process involved in a reporting environmental management information system.

Custom alarm identification and formal alarm calculation are simple tasks, for which environmental experts or legislation have specified, respectively, well-defined, explainable rules. In this respect, it is a task easy to be automated and reproduced by a computer system. However, the automated decision-making process, which involves the *validation of incoming measurements* and *the estimation of missing ones*, is a challenging problem, which is in principal dependent on local conditions and seasonal trends.

4. Mining Air Quality Data

4.1. Knowledge Discovery for Air Quality Assessment

The issues raised in an operational air quality assessment framework may be tackled using knowledge discovery techniques. Our assumption is that mining air quality data may yield to trustworthy predictive models, which can be used for supporting future decision-making. Interesting patterns, hidden in environmental data sets, can be discovered and subsequently embedded into an EMIS operational decision support module. In this way, data-driven decision making models can be used for supporting an automated procedure for issuing air quality alarms at a timely fashion. Such an EMIS can be realized in a decision support system for assessing ambient air quality, as the developed multi-agent system *O_3RTAA* (Athanasiadis and Mitkas, 2004a, b).

A data-driven solution for dealing with uncertainties in an environmental monitoring network is preferable, because it takes into account the local characteristics of the problem at hand, which may deviate from general trends or 'rules of thumb'. As a result, the overall decision-making is more accurate, since data-driven models adapt the problem-solving method to local conditions and time-evolving trends.

4.2. Available Data and Preprocessing

The ability of knowledge discovery techniques to deal with data uncertainty problems involved in an EMIS is de-

Table 1. Air Quality Indicators

Pollutant		Air Quality Bands						
Name	Units	Low	S	Moderate	I	High	A	Very High
Ground Ozone	ppb (1 h av.)	< 50		50-89		90-179		> 180
Carbon Monoxide	ppb (8 h r.av.)	< 10		10-14		15-19		> 20
Nitrogen Dioxide	ppb (1 h av.)	< 150		150-299		300-399		> 400
Sulphur Dioxide	ppb (15 min av.)	< 100		100-199		200-399		> 400

Note: S - Standard threshold, I - Information Threshold, A - Alerting Threshold.

monstrated in the followings. Specifically, two decision models (estimators) have been developed: one to validate the incoming ozone measurements and another to estimate the missing ones. These estimators realize data-driven strategies induced from environmental data recorded by an air-quality monitoring network in the district of Valencia, Spain.

The available data originated from three meteorological stations, situated in distinct locations in the region of Valencia. Nine attributes, including both meteorological and air-pollutant variables, were sampled on a quarter-hourly basis for a period covering years 1999 to 2001. The sampled variables as well as their corresponding units are shown in Table 2. The recorded measurements are accompanied with the respective validation tags and quality indicators for ambient ozone variable, which have been appended manually by environmental scientists. The ozone validation tag characterizes the corresponding ozone measurement as correct ('a') or erroneous ('l'). The ozone concentration level is characterized as either 'low', 'medium', 'high' or 'very high' for values in the ranges 0 ~ 49, 50 ~ 89, 90 ~ 179, or more than 180 μg/m^3, respectively. These ranges correspond to the 'Air Quality Bands' of the Valencian Community (Mantilla and Perez, 2002).

Datasets contain 105,216 records for each station, bringing the total to 315,648 data records. In 16,304 records, that is around 5.2% of the total, the ozone variable is characterized as 'erroneous'. Errors in measuring ozone concentration may be attributed to several reasons, including polarization, noise, network or sensor fault. For more than half of the erroneous records, the ozone concentration is missing, while for the rest some measurement is recorded, but it was rejected by the environmental scientists. Ozone air quality indicator is classified in four labels: 'L', 'M', 'H', and 'V', with the overall distribution, in all datasets, at 27.8, 41.7, 27.5 and 0.2%, respectively. The statistics of the nine available datasets are presented in Table 3.

4.3. The Measurement Validation Predictive Model

Measurement validation is in principle a function approximation problem, typically addressed in sensor networks using statistical methods (Cox et al., 2000), principal component analysis (Harkat et al., 2004), Kalman filters (Schneider and Oezguener, 1998), belief networks (Alag et al., 2001), or association rules (Yairi et al., 2001). However, the problem at hand is to decide whether an ozone measurement captured by the sensor is valid or not. As there are two validation tags (namely 'a' and 'l') the incoming measurement validation problem essentially becomes a two-class classification problem. Due to the uneven distribution of the two classes it is essential to focus on the identification of the minority class ('l').

Table 2. Air Pollutants and Meteorological Attributes

Data Attribute		Sym.	Data Type	Units
1	Date	D	date	
2	Time	T	time	
3	Sulfur dioxide	SO_2	real	g/m^3
4	Ozone	O_3	real	g/m^3
5	Nitrogen oxide	NO	real	g/m^3
6	Nitrogen dioxide	NO_2	real	g/m^3
7	Nitrogen oxides	NO_x	real	g/m^3
8	Wind velocity	VEL	real	m/s
9	Wind direction	DIR	real	deg
10	Temperature	TEM	real	^{o}C
11	Relative humidity	HR	real	%
12	O_3ValidationTag	VAL	categorical	'a' correct 'l' erroneous
13	Ozone Indicator	O_3 Level	categorical	'L' (0 - 49 g/m^3) 'M' (50 - 89 g/m^3) 'H' (90 - 179 g/m^3) 'V' (> 180 g/m^3)

The predictive model developed to estimate the validity of an incoming ozone measurement uses the immediate history of the ozone sensor recordings:

(1) the current value for which the predictor has to decide its validity,

(2) three prior ozone measurements among those observed within a time window of one and a half hour, and

(3) three measures of ozone value variation in the past ninety minutes.

The available time-series data from all nine datasets have been preprocessed and the extracted features are presented in Table 4. The predictive model comprises seven predictor variables, calculated within a time window of 90 minutes (i.e. the past six measurements are buffered), and the response vari-

Table 3. Environmental Datasets Statistics

				Ozone Measurements		Ozone Quality Indicator			
Dataset No.	Station	Year	Instances	Valid	Erroneous	L	M	H	V
1		1999	35040	33390	1650	15931	12366	5405	88
2	GRAU	2000	35136	33699	1437	17878	11109	4830	120
3		2001	35040	33187	1853	19971	12294	1742	108
4		1999	35040	30470	4569	1082	14570	16240	46
5	MORE	2000	35136	31881	3255	2653	15054	16779	43
6		2001	35040	33041	1998	2575	15116	16068	24
7		1999	35040	34318	722	8626	17279	8851	46
8	ONDA	2000	35136	34881	255	9241	17460	8267	21
9		2001	35040	34475	565	9777	16283	8482	29

able 'O3val', which is a nominal attribute labeled 'a' or 'l'.

4.4. Erroneous Measurement Estimation Predictive Model

An estimation of ambient ozone concentration from other variables is "feasible in principle". The ambient ozone concentration is known to be a function of both nitrogen oxides NO_x (Clappa and Jenkin, 2001) and meteorological variables (Huang and Smith, 1997). It has been demonstrated that estimation of environmental missing data can be affected by regression techniques, e.g. linear extrapolation (Hedges, 1999). Nevertheless, conventional repressor models are restricted by a priori assumptions, including a model's structure. An empirical approach is proposed here for estimating missing ozone measurements directly from the data 'by classification'.

Table 4. Attributes Used for the Validation Decision Model

O3	The current ozone value
O3_15	The ozone value 15 min ago
O3_45	The ozone value 45 min ago
O3_75	The ozone value 75 min ago
MinMax30	The difference between the maximum and the minimum ozone value in the last 30 minutes
MinMax60	The difference between the maximum and the minimum ozone value in the last 60 min
MinMax90	The difference between the maximum and the minimum ozone value in the last 90 min
O3val	The corresponding validation tag (valid/erroneous)

The problem can be summarized as follows: when the ozone sensor captures no measurement, or if the captured measurement is rejected by the validation process, the goal is to estimate the missing ozone concentration value from the remaining variables available. This problem can be considered as a function approximation problem. Since there are only four ozone concentration levels, the aforementioned estimation problem can be reformed as a classification problem.

Specifically, we have developed two predictive models for estimating the ambient ozone's concentration levels. The first one uses only the concurrent measurements of other pollutants and meteorological attributes for predictor variables. In this way, an on-line, memoryless, decision-making scheme is created, as only concurrent measurements are used. In the second model, historical ozone measurements are appended. This model uses a short memory for storing past ozone measurements in a 30 minutes buffer, i.e. the prior two measurements are cached. The output variable in both models is the ozone quality indicator, a nominal variable sized 'L', 'M', 'H', 'V'. The attributes used for these models are summarized in Table 5. The available datasets have been preprocessed properly in order to be restructured in the appropriate form.

Table 5. Attributes Used for the Estimation Decision Model (Those in italics are used in the second model)

SO_2	The concurrent value of SO_2 concentration
NO	The concurrent value of NO concentration
NO_2	The concurrent value of NO_2 concentration
NO_X	The concurrent value of NO_x concentration
VEL	The concurrent value of Wind velocity
TEM	The concurrent value of Temperature
HR	The concurrent value of Relative Humidity
O_3_15	The ozone value 15 min ago
O_3_30	The ozone value 30 min ago
O_3Class	The (missing) ozone value level (low/med

5. Model Evaluation and Results

5.1. Decision Tree Induction

An empirical approach for creating data-driven decision making models was utilized for both cases. Quinlan's C4.5 algorithm for decision tree induction was employed (Quinlan, 1991), as one of the most widely-used, state-of-the-art classifier. Specifically, the C4.5 implementation in WEKA knowledge analysis environment (Witten and Frank, 1999), named J48, was used. The J48 has been employed for inducing both pruned and un-pruned decision trees, whose nodes specify inequalities for the values of the respective environmental pre-

Table 6. Measurement Validation Model Training and Testing

Dataset		Training Phase			Testing Phase		
Station	Year	Percent correct	Scheme options	Number of rules	Overall accuracy	Erroneous Measurements	
						Precision	Recall
GRAU	1999	99.62	N 5	10	98.64	89.47	40.91
	2000	99.77	N 3	4	99.68	98.88	83.02
	2001	99.91	U	35	95.29	93.71	57.20
MORE	1999	99.69	C 0.5	13	97.45	96.92	85.41
	2000	98.41	N 30	15	99.47	88.12	97.67
	2001	98.69	N 10	15	97.48	86.79	79.96
ONDA	1999	99.64	N 3	3	97.72	93.03	32.69
	2000	99.75	C 0.05	8	99.80	61.90	78.00
	2001	99.86	N 3	4	99.74	99.69	88.01

dictor attributes, while its leaves specify the output class. Two pruning methods have been used for improving the decision making capabilities of the induced decision trees: a. Confidence Factor Pruning, and b. Reduced Error Pruning.

In total, we evaluated twenty-three training schemes with C4.5 algorithm, using the following options:

(1) Un-pruned decision tree induction (U). (One scheme)

(2) Pruned decision tree induction, using Confidence Factor parameter (C), where C = 0.05, 0.1, …, 0.45, 0.5 (10 schemes).

(3) Pruned decision tree induction, with Reduced Error Prunning using various values for the number of folds parameter (N), where N = 2, 3, 5, 10, 20, …, 500, 1000 (12 schemes).

5.2. Training and Testing

The aforementioned training schemes have been used for inducing decision trees for both the incoming measurement validation and the erroneous measurement estimation tasks. The twenty-three training schemes have been applied on all preprocessed datasets. Training has been performed independently for each station and each year, i.e. there have been nine set of experiments for each case. A uniform training and testing procedure was followed: For each experiment, the first half of the records, covering period January to June of the year, has been used for training. The remaining records, which correspond to the period July to December of the year, have been used for testing. Following this procedure, C4.5 capability for learning from data is investigated for creating data-driven decision making models that are adapted in both space (i.e. station) and time (i.e. year). In total, we elaborated $9 \times 23 = 207$ experiments for each predictor.

5.3. Results for the Incoming Measurement Validation Task

On overview of the results acquired for the incoming measurement validation task, is shown in Table 6. The results of the decision tree that outperforms for each experiment are presented along with the scheme options and the number of rules for the training phase. The overall accuracy at the testing phase in all experiments is extremely satisfactory, as its average reaches 98.3%. As the minority class corresponds to the 5.2% of the total records, we consider 95% accuracy performance as a "measure of acceptance" for the induced models. In this respect, we consider the extracted decision trees markedly capable of validating incoming ozone measurements. Also, note that the minority class is correctly identified. Minority class identification precision is over 88% in most cases, while minority class recall measure reaches an average of 71.5%.

Table 7. Erroneous Measurement Estimation Model Results (Model without history)

Dataset		Training Accuracy	Scheme Options	Number of Rules	Testing Accuracy
Station	Year				
GRAU	1999	79.45	N 300	53	74.29
	2000	79.66	N 300	40	74.91
	2001	86.15	N 200	72	75.48
MORE	1999	87.41	N 20	248	67.99
	2000	87.29	N 2	355	59.54
	2001	84.52	N 30	235	52.72
ONDA	1999	75.37	N 300	63	62.93
	2000	73.22	N 200	85	61.69
	2001	65.72	N 1000	24	57.68

5.4. Erroneous Measurement Estimation Results

As presented in section 4.3, two predictive models have been developed for estimating the missing or erroneous ozone concentration levels. The first one uses concurrent pollutant and meteorological variable values. The corresponding results are shown in Table 7. The second combines concurrent pollutant and meteorological variable values with ozone's immediate history. Its results are summarized in Table 8. The memoryless models have a very good performance, but the mo-

dels with history outperform in all cases. Outstanding results, acquired with the models with history, have an average predictive accuracy of 93.75% on the test data. Also, note that decision trees induced for the history model are simpler, as the number of rules is smaller. Experimental results imply that knowledge discovery techniques can produce to create decision making models that estimate ozone's erroneous measurements successfully.

Table 8. Erroneous Measurement Estimation Model Results (Model with history)

Dataset		Training Accuracy	Scheme Options	Number of Rules	Testing Accuracy
Station	Year				
	1999	94.03	N 10	71	94.23
GRAU	2000	94.97	C 0.05	68	94.63
	2001	96.29	N 2	68	93.32
	1999	97.58	N 50	7	95.98
MORE	2000	96.69	C 0.05	4	97.03
	2001	97.49	N 50	20	95.57
	1999	90.68	N 1000	3	90.66
ONDA	2000	91.70	N 3	64	91.86
	2001	92.45	N 30	52	90.49

6. Conclusions

In this paper, knowledge discovery techniques have been applied for supporting the decision-making process involved in an operational assessment module reporting EMIS. While previous work has been concentrated in forecasting problems, in this paper we tackled issues related to on-line decision-making and operational air quality assessment. The empirical approach followed yielded trustworthy decision making models, as analytically shown in the previous section. The results of this study have brought forth the potential value of data-driven approaches for operational air quality assessment and decision making. Both issues related to measurement validation and missing measurement estimation levels have been tackled efficiently using data mining techniques. In the case of validating ozone measurements a decision model with a short memory (of ninety minutes in our experiments) was sufficient in most data sets for ensuing reliable decisions in most cases. Our empirical results shown that in average only a portion around ten per cent of the invalid measurements were not identified as such. The accuracy rate achieved is acceptable for an automated system that can review the incoming data streams with no supervision. On the second front of estimating the missing or erroneous measurements, this paper introduced two qualitative estimators instead of the common approach for quantitative predictors. When a measurement is missing or is erroneous, instead of employing a regression estimator, that tries to estimate a numerical value, in this study we employed a classifier, that estimates the missing measurement level, according to the specified air quality bands. A more detailed comparative analysis between statistical methods and classification algorithms for air quality forecasting was presented in Athanasiadis et al. (2005). The application of classification methods for operational air quality assessment is a novelty introduced here, which produced trustworthy models with predictive accuracy that exceeded 93%. To conclude with, data-driven, classification approaches managed to deal with data uncertainties involved in an air quality EMIS, in order to support decision makings in an operational time frame.

Acknowledgments. Authors would like to express their gratitude to the IDI-EIKON team for their efforts within Agent Academy project to deploy the O3RTAA system and to CEAM for the provision of the environmental datasets. The Agent Academy project was partially funded by the European Commission under the IST programme (IST-2000-31050).

References

Alag, S., Agogino, A.M. and Morjaria, M. (2001). A methodology for intelligent sensor measurement, validation, fusion, and fault detection for equipment monitoring and diagnostics. *Artif. Intell. Eng. Des., Anal. Manuf.*, 15, 307-320.

Athanasiadis, I.N., Karatzas, K. and Mitkas, P.A. (2005). Contemporary air quality forecasting methods: A comparative analysis between statistical methods and classification algorithms, in R. Sokhi and J. Brexhler (Ed.), *5th Int'l Conference on Urban Air Quality Measurement, Modelling and Management*.

Athanasiadis, I.N. and Mitkas, P.A. (2004a). An agent-based intelligent environmental monitoring system. *Manag. Environ. Qual.*, 15, 238-249.

Athanasiadis, I.N. and Mitkas, P.A. (2004b). Applying agent technology in environmental management systems under real-time constraints, in C. Pahl, S. Schmidt, A.E. Rizzoli and A. Jakeman (Ed.), *Second Biennial Meeting of the Int'l Environmental Modelling and Software Society: Complexity and Integrated Resources Management*, 2, pp. 531-536.

Athanasiadis, I.N., Kaburlasos, V.G., Mitkas, P.A. and Petridis, V. (2003). Applying machine learning techniques on air quality data for real-time decision support, *First Int'l Symposium on Information Technologies in Environmental Engineering* (ITEE-2003), ICSC-NAISO Academic Press, pp. 51.

Athanasiadis, I.N., Mitkas, P.A., Laleci, G.B. and Kabak, Y. (2003). Embedding data-driven decision strategies on software agents: The case of a multi-agent system for monitoring air-quality indexes. in R. Jardim-Goncalves, J. Cha and A. Steiger-Garcao (Ed.), *Concurrent Engineering: The Vision for the Future Generation in Research and Applications*, Balkema Publishers, 1, pp. 23-30.

Barratt, B. (2000). *The Hertfordshire and Bedfordshire Air Pollution Monitoring Network*, SEIPH-Environmental Research Group, King's College, UK.

Bordignon, S., Gaetan, C. and Lisi, F. (2002). Nonlinear models for ground-level ozone forecasting. *Stat. Methods Appl.*, 11, 227-246.

Chen, L., Islam, S. and Biswas, P. (1998). Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction. *Atmos. Environ.*, 32, 1839-1848.

Clappa, L.J. and Jenkin, M.E. (2001). Analysis of the relationship between ambient levels of O3, NO2 and NO as a function of NOX in the UK. *Atmos. Environ.*, 35, 6391-6405.

Cox, M.G., Harris, P.M., Milton, M.J.T. and Woods, P.T. (2002). *Method for Evaluating Trends in Ozone Concentration Data and its Application to Data from the UK Rural Ozone Monito-*

ring Network, National Physical Laboratory, Crown Publishers, UK.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases (a survey). *AI Mag.*, 3(17), 37-54.

Gibert, K., Spate, J., Sánchez-Marré, M., Frank, E., Comas, J. and Athanasiadis, I. (2007). iEMSs 2006 Position Papers. In A. Voinov, A. Jakeman, and A.E. Rizzoli (Ed.), *Data Mining for Environmental Systems* (to appear).

Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques*, Morgan-Kaufmann.

Harkat, M., Mourot, G. and Ragot, J. (2004). Sensor failure detection and isolation of an air quality monitoring network using Principal Component Analysis, *Proc. of the Symposium Techniques Avancées et Stratégies Innovantes en Modélisation et Commandes Robustes des Processus Industriels*, ISA, The Instrumentation, Systems and Automation Society.

Hedges, S. (1999). *Ozone Monitoring, Mapping, and Public Outreach: Delivering Real-Time Ozone Information to Your Community*, US Environmental Protection Agency, Cincinnati, Ohio, USA.

Huang, G.H. and Chang, N.B. (2003). Perspectives of Environmental Informatics and Systems Analysis. *J. Environ. Inf.*, 1, 1-6.

Huang, L. and Smith, R.L. (1999). Meteorologically-dependent Trends in Urban Ozone. *Environmetrics*, 10, 103-118.

Kalapanidas, E. and Avouris, N. (2002). Air quality management using a multi-agent system. *Int. J. Comput. Aided Civil Infrastruct. Eng.*, 17, 119-130.

Kim, H. and Guldmann, J. (2001). Modeling air quality in urban areas: A cell-based statistical approach. *Geogr. Anal.*, 33, 156-180.

Klosgen, W., Zytkow, J.M. and Zyt, J. (2002). *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press.

Larssen, S. and Hagen, L.O. (1996). *Air pollution monitoring in Europe. Problems and Trends*, European Topic Centre on Air Quality, European Environment Agency, Copenhagen, Denmark.

LAQN (The London Air Quality Network). http://www.erg.kcl.ac.uk/London.

Lekkas, G.P., Avouris, N.M. and Viras, L.G. (1994). Case-based reasoning in environmental monitoring applications. *Appl. Artif. Intell.*, 8, 359-376.

Mantilla E. and Perez J.G. (2002). Strategies for the automated evaluation of meteorological temporary series and air quality. CEAM Report, Fundacion Centro de Estudios Ambientales del Mediteraneo, Spain.

Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning*, Morgan Kaufmann.

Ruiz-Suárez, J.C., Ibara, O.A.M., Jiménez, J.T. and Suárez, L.G.R. (1995). Short-term ozone forecasting by artificial neural networks. *Adv. Eng. Softw.*, 23, 143-149.

Schneider, S.J. and Oezguener, U. (1998). A framework for data validation and fusion, and fault detection and isolation for intelligent vehicle systems, *Proc. of the IEEE International Conference on Intelligent Vehicle*s, pp. 533-538.

TCEQ (The Texas Commission on Environmental Quality). http://www.tceq.state.tx.us.

Witten, I.H. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

Yairi, T., Kato, Y. and Hori, K. (2001). Fault Detection by Mining Association Rules from House-keeping Data, *Proc. of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space*.

Yi, J. and Prybutok, V.R. (1996). A neural network model for the prediction of daily maximum ozone concentration in an Industrialized urban area. *Environ. Pollut.*, 92, 349-357.