In **Knowledge Discovery for Environmental Management**, (Voss, H., Wachowicz, M., Dzeroski, S., and Lanza, A., Eds.), Bonn, Germany, KDnet, June 2004, pp.1-12.

Supporting the Decision-Making Process in Environmental Monitoring Systems with Knowledge Discovery Techniques

Ioannis N. Athanasiadis and Pericles A. Mitkas

Informatics and Telematics Institute, Centre for Research and Technology Hellas
and
Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki, Greece
ionathan@ee.auth.gr, mitkas@eng.auth.gr

Abstract. In this paper an empirical approach for supporting the decision making process involved in an Environmental Management System (EMS) that monitors air quality and triggers air quality alerts is presented. Data uncertainty problems associated with an air quality monitoring network, such as measurement validation and estimation of missing or erroneous values, are addressed through the exploitation of data mining techniques. Exhaustive experiments with real world data have produced trustworthy predictive models, capable of supporting the decision-making process. The outstanding performance of the induced predictive models indicate the added value of this approach for supporting the decision making process in an EMS.

1. Introduction

1.1. Environmental monitoring

Environmental monitoring networks have been established worldwide in order to observe the conditions of the natural environment. Such networks generate vast volumes of raw data, while information systems, called Environmental Management Systems (EMS), have been occupied with integrating all recorded data-streams. A typical EMS installation involves the fusion into a central database of all data sensed at distributed locations. Until lately, all recorded data were meant for environmental scientists 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 few years 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, near-real–time (NRT) environmental assessment, incident identification and reporting services have to be incorporated in EMS. Such services require decision making at 'near real time'. The NRT constraint reveals two critical problems in delivering such tasks: (a) the low quality or absence of data, and (b) the changing conditions over a long period of time. As a result, the critical properties of an automated decision-making system are its ability to validate incoming measurements and its ability to adapt to an ever-changing environment. 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. Conditions for environmental management keep changing with time, demanding periodically updated decision support [14]. These properties can be realized by *learning from data*, using knowledge discovery techniques.

1.2. Predictive models induced from environmental monitoring data

Earlier research work has dealt with EMS using knowledge discovery techniques mainly for incident forecasting. Several models have been built for predicting incidents that may occur in the near future. For instance, conventional statistical regression models [7,15,17] and time-series analysis have been applied to predict ozone levels [8]. Neural networks have been used for short-term ozone prediction [22,26], while case-based reasoning [19] and classification and regression trees [16] have been employed for predicting air pollutant concentrations. Another example is the system developed for generating coral bleaching alerts, which is an indicator of harsh environmental conditions in an aquatic ecosystem, using a data-driven expert system [13]. 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.

In this work, knowledge discovery techniques have been applied for supporting decision making processes involved in an EMS, 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 real-time and NRT reporting systems. Decision support is also necessary for distributing trustworthy information with minimal human intervention at the time incidents occur. In this paper, we present our work conducted for the development of data-driven decision strategies for successfully assessing ambient air quality at 'near real time', through the exploitation of machine learning techniques.

2. Ambient air quality assessment

2.1. Domain background

Air quality depreciates in many cities, as a result of industrial activities and traffic emissions. For this reason, Air Quality Operational Centers have established monitoring networks in areas with (potential) air pollution problems. These networks sense atmospheric conditions and trace related measurements, such as meteorological attributes and pollutants 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 [6]. EU Directives on Air Quality have delimited '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.

Table 1. Air quality indicators

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

S: Standard threshold, **I** : Information Threshold, **A** : Alerting Threshold

In environmental monitoring networks, various 'sensor breakdown events', such as sensor malfunction, network delay, or noise, may lead to loss 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.

In this manner, 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 [12]. In London Air Quality Network[1], flexible data analysis is supported through statistical tools [6] and in Texas Natural Resource Conservation Commission[2], meteorologists set criteria for validating data and making predictions.

It is evident that even if the calculation of air quality indicators is a simple task, the preparatory activities, of data validation and missing measurement estimation are complex processes, typically undertaken by humans. Even if the recorded measurements are available at time incidents occur, data validation and review is a

[1] http://www.erg.kcl.ac.uk/london
[2] http://www.tnrcc.state.tx.us

struggling task that makes the whole procedure time-consuming. The European Environment Agency indicates that validated data, reviewed by environmentalists, are available one to six months after measurement [18]. Therefore, automating this procedure with respect to near real-time constraints is valuable.

2.2. The related decision-making process

Two are the driving forces involved in ambient air quality assessment. First comes the societal need for information, dispersed at the time of an incident. The second is data uncertainty, which is translated into an enormous workload for environmental scientists. In this context, an EMS 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.

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, dependent on local conditions and seasonal trends. This problem was tackled in this work using knowledge discovery techniques on environmental monitoring data.

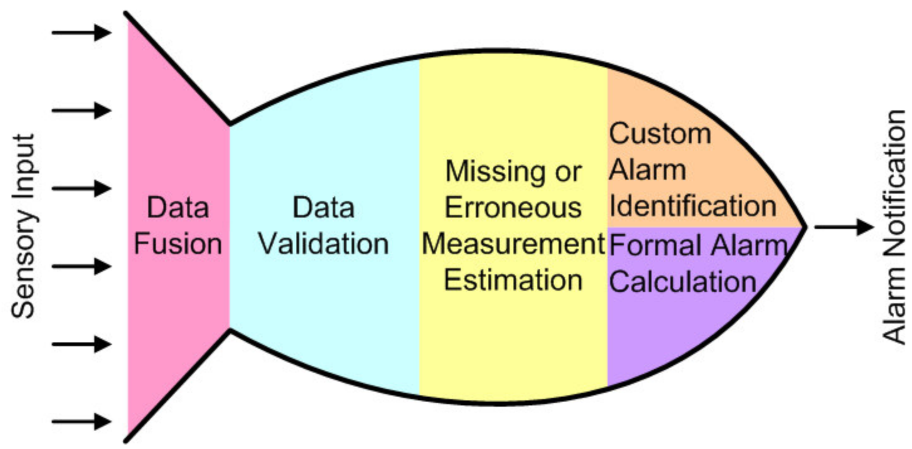

Fig. 1. The decision making process involved in reporting EMS.

2.3. Knowledge discovery for air quality assessment

The current procedure at Air Quality Operational Centers involves data validation and missing measurement estimation by human experts. Vast volumes of data recorded by monitoring networks have been reviewed by experts, who appended data validation tags and air quality indicators. Our assumption is that mining these data may yield 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 a decision support system. 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. 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.

3. Mining Air Quality Data

3.1. Available data and preprocessing

In this work, we demonstrate the ability of knowledge discovery techniques to deal with data uncertainty problems involved in an EMS. Specifically, two estimators are employed to validate incoming ozone measurements and to estimate 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 had been collected in three meteorological stations, situated in distinct locations in the region of Valencia. Nine variables, including both meteorological and air-pollutant variables, were sampled on a quarter-hourly basis for a period covering years 1999-2001. 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 apended 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 $\mu g/m^3$, 50–89 $\mu g/m^3$, 90–179 $\mu g/m^3$, or 180– $\mu g/m^3$, respectively. These ranges correspond to the 'Air Quality Bands' of the Valencian Community [20].

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, 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.

Table 2: Air pollutants and meteorological attributes

Data	Attribute	Symbol	Data Type	Units
1	Date	D	date	
2	Time	T	time	
3	Sulfur dioxide	SO_2	real	$\mu g/m^3$
4	Ozone	O_3	real	$\mu g/m^3$
5	Nitrogen oxide	NO	real	$\mu g/m^3$
6	Nitrogen dioxide	NO_2	real	$\mu g/m^3$
7	Nitrogen oxides	NO_x	real	$\mu g/m^3$
8	Wind velocity	VEL	real	m/s
9	Wind direction	DIR	real	deg
10	Temperature	TEM	real	^oC
11	Relative humidity	HR	real	%
12	O_3ValidationTag	VAL	nominal	'a' correct
				'l' erroneous
13	Ozone Indicator	O_3Level	nominal	'L' (0-49$\mu g/m^3$)
				'M' (50-89$\mu g/m^3$)
				'H' (90-179 $\mu g/m^3$)
				'V' (>180$\mu g/m^3$)

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

3.2. Incoming measurement validation predictive model

Measurement validation is, in general, a function approximation problem, typically addressed in sensor networks using statistical methods [10], principal component analysis [11], Kalman filters [23], belief networks [1], or association rules [25]. 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’).

The predictive model developed to estimate the validity of an incoming ozone measurement uses the immediate history of the ozone sensor recordings. Ozone time-series data recorded in the nine available datasets have been preprocessed and the extracted features are described in Table 4.

The predictive model comprises seven predictor variables, calculated for a time window of 90 minutes (i.e. the past six measurements are buffered), and the response variable 'O3val', which is a nominal attribute labeled 'a' or 'l'.

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)

3.3. Erroneous measurement estimation predictive model

An estimation of ambient ozone concentration from other variables is “feasible in principle”. Ambient ozone concentration is known to be a function of both nitrogen oxides NO_x [9] and meteorological variables [15]. It has been demonstrated that estimation of environmental missing data can be affected by regression techniques, e.g. linear extrapolation [12]. 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'.

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 ambient ozone's concentration level. 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)

4. Experiments and results

4.1. 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 [21]. Specifically, the C4.5 implementation in WEKA knowledge analysis environment [24], named J48, was used. J48 has been employed for inducing both pruned and un-pruned decision trees, whose nodes specify inequalities for the values of the respective environmental predictor 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 tested twenty three training schemes with C4.5 algorithm, using the following options:

a) Un-pruned decision tree induction (U). (One scheme)
b) Pruned decision tree induction, using the Confidence Factor parameter (C), where C = 0.05, 0.1, …, 0.45, 0.5 (10 schemes).
c) 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).

4.2. Training and testing

The aforementioned training schemes have been used for inducing decision trees for both the incoming measurement validation task and the erroneous measurement estimation task. 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 experiments for each case. A uniform training and testing procedure was followed: For each experiment, the first half of the records, covering period January – June of the year, has been used for training. The remaining records, which correspond to the period July–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 9×23=207 experiments for each task.

4.3. 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 6: Measurement Validation Model Training and Testing

		Training Phase			**Testing Phase**		
Dataset		*Percent*	*Scheme*	*Number*	*Overall*	*Erroneous Measurements*	
Station	*Year*	*Correct*	*Options*	*of Rules*	*Accuracy*	*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

4.4. Erroneous measurement estimation results

As discussed in section 3.3, two predictive models have been developed for estimating 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 models 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 7: Erroneous Measurement Estimation Model Results
(Model without history)

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

Table 6: Erroneous Measurement Estimation Model Results
(Model with history)

Dataset		Training	Scheme	Number	Testing
Station	Year	Accuracy	Options	of Rules	Accuracy
	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

5. Discussion

In this paper, knowledge discovery techniques have been applied for supporting the decision-making process involved in a reporting EMS. The empirical approach followed yielded trustworthy decision making models. Experimental results have proven the potential of knowledge discovery techniques for data-driven, on-line decision support, which adapts to local conditions and time-evolving trends. Specifically, the data-driven approach, managed to deal with data uncertainties involved in an air quality EMS.

The predictive models extracted have been realized in a NRT reporting EMS, named O_3RTAA [2,3]. The O_3RTAA system has been developed as a multi-agent decision support system for assessing ambient air quality. O_3RTAA system has been deployed in collaboration with IDI-EIKON, Valencia, Spain, and has been successfully installed at the Mediterranean Centre for Environmental Studies Foundation (CEAM), Valencia, Spain.

The overall goal of O_3RTAA is to assess air quality by the identification of ambient ozone indicators. O_3RTAA captures air-quality data, including pollutants' concentrations, measured at several meteorological stations and processes them in order to calculate both Formal and Custom Ozone Alarms. Besides the alarm triggering activities and the usual 'housekeeping' tasks, such as the updating of the database with incoming measurements, O_3RTAA is empowered [4,5] with advanced features, including measurement validation, and estimation of missing values. These advanced features are enabled by agent capabilities for reasoning rationally and the exploitation of knowledge discovery techniques.

Acknowledgements

Authors would like to express their gratitude to the IDI-EIKON team for their efforts within Agent Academy project to deploy the O_3RTAA system and to CEAM for the provision of the environmental datasets. The Agent Academy project is partially funded by the European Commission under the IST programme (IST-2000-31050).

References

1. S. Alag, A. M. Agogino, and M. Morjaria. A Methodology For Intelligent Sensor Measurement, Validation, Fusion, And Fault Detection For Equipment Monitoring And Diagnostics, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 15(4):307-320, 2001.
2. I. N. Athanasiadis and P. A. Mitkas. An agent-based intelligent environmental monitoring system. *Management of Environmental Quality*, 15 (3), 2004.
3. I. N. Athanasiadis and P. A. Mitkas. Applying agent technology in environmental management systems under real-time constraints. In *Proceedings of the Second Biennial Meeting of the International Environmental Modelling and Software Society*, Osnabrueck, Germany, 2004. iEMSs.
4. I. N. Athanasiadis, V. G. Kaburlasos, P. A. Mitkas, and V. Petridis. Applying machine learning techniques on air quality data for real-time decision support. In Information Technologies in Environmental Engineering, Gdansk, Poland, ICSC-NAISO Publishers, June 2003.
5. I. N. Athanasiadis, P. A. Mitkas, G. B. Laleci, and Y. Kabak. Embedding data-driven decision strategies on software agents: The case of a multi-agent system for monitoring air-quality indexes. In Concurrent Engineering: The Vision for the Future Generation in Research and Applications, Madeira, Portugal, Balkema Publishers, July 2003.
6. B. Barratt. The Hertfordshire and Bedfordshire air pollution monitoring network. Annual Report 1999, SEIPH - Environmental Research Group, King's College, London, UK, 2000.
7. S. Bordignon, C. Gaetan and F. Lisi. Nonlinear models for ground-level ozone forecasting. Statistical Methods and Applications, 11:227-246, 2002.
8. L.-J. Chen, S. Islam, and P. Biswas. Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction. *Atmospheric Environment*, 32(11):1839-1848, 1998.
9. L. J. Clappa and M. E. Jenkin. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NO_X in the UK. *Atmospheric Environment*, 35:6391–6405, 2001.

10. M. G. Cox, P. M. Harris, M. J. T. Milton and P. T. Woods. Method for Evaluating Trends in Ozone Concentration Data and its Application to Data from the UK Rural Ozone Monitoring Network, NPL Report CMSC 15/02, National Physical Laboratory, Middlesex, UK, Crown Publishers, November 2002
11. M.-F. Harkat, G. Mourot, J. Ragot. Sensor failure detection and isolation of an air quality monitoring network using Principal Component Analysis. In Proceedings of the Symposium Techniques Avancées et Stratégies Innovantes en Modélisation et Commandes Robustes des Processus Industriels, 2004.
12. S. Hedges. Ozone monitoring, mapping, and public outreach: Delivering real-time ozone information to your community. Technical Report EPA-625-R-99-007, United States Environmental Protection Agency, Cincinnati, Ohio, USA, September 1999.
13. J. C. Hendee, E. Mueller, C. Humphrey, and T. Moore. A data-driven expert system for producing coral bleaching alerts at Sombrero Reef in the Florida Keys. *Bulletin of Marine Science*, 69(2):673-684, 2001.
14. G.H. Huang and N. B. Chang. Perspectives of environmental informatics and systems analysis. *Journal of Environmental Informatics*, 1(1):1-6, 2003.
15. L. S. Huang and R. L. Smith. Meteorologically-dependent trends in urban ozone. Technical Report 72, National Institute of Statistical Sciences, 1997.
16. E. Kalapanidas and N. Avouris. Short-term air quality prediction using a case based classifer. *Environmental Modelling and Software*, 16(3):263-272, 2001.
17. H.-Y. Kim and J.-M. Guldmann. Modeling air quality in urban areas: A cell-based statistical approach. *Geographical Analysis*, 33, 2001.
18. S. Larssen and L. O. Hagen. Air pollution monitoring in Europe. Problems and trends. Topic Report 26-1996, European Topic Centre on Air Quality, European Environment Agency, Copenhagen, Denmark, November 1996.
19. G. P. Lekkas, N. M. Avouris, and L.G. Viras. Case-based reasoning in environmental monitoring applications. *Applied Artificial Intelligence*, 8:359-376, 1994.
20. E. Mantilla. and J. G. Perez. Strategies for the automated evaluation of meteorological temporary series and air quality. Ceam Report, Fundacion Centro de Estudios Ambientales del Mediteraneo, September, 2002.
21. J.R.Quinlan. *C4.5 Programs for Machine Learning*, Morgan Kaufmann, 1993.
22. J. C. Ruiz-Suarez, O. A. Mayora-Ibara, J. Torres-Jimenez, and L. G. Ruiz-Suarez. Short-term ozone forecasting by artificial neural networks. *Advances in Engineering Software*, 23:143-149, 1995.
23. S. J. Schneider,and U. Oezguener. A Framework for Data Validation and Fusion, and Fault Detection and Isolation for Intelligent Vehicle Systems. In *Proceedings of the IEEE International Conference on Intelligent Vehicles*, pp. 533-538, IEEE, 1998.
24. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.
25. T. Yairi, Y. Kato and K. Hori. Fault Detection by Mining Association Rules from House-keeping Data. In *Proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, Montreal, June 2001.
26. J. Yi and V. R. Prybutok. A neural network model for the prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92:349-357, 1996.