In **Managing Environmental Knowledge (Proc. 20th Int'l Conference on Informatics for Environmental Protection: EnviroInfo 2006)**, (Tochtermann, Klaus and Scharl, Arno, Eds.), Graz, Austria, Shaker Verlag, September 2006, pp.505-508.

Data mining air quality data for Athens, Greece

Melina Efraimidou[1], Maria Kanaki[1], Ioannis Athanasiadis[2],
Pericles Mitkas[3] and Kostas Karatzas[4]

Abstract

Urban air quality management and information systems are more frequently required to include advanced capabilities of quick, effective and easy to operate environmental data analysis applications for information extractions and the support of decision making. These systems are based on the need of city authorities and national governments to establish a framework which enables them to take actions, in order to ensure that air quality is improved and relevant standards are maintained in urban areas. 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. In addition, conditions for environmental management keep changing with time, demanding periodically updated decision support. These properties can be realized by learning from data, using knowledge discovery techniques. In the present paper, data mining techniques are applied for data analysis and for the construction of forecasting modules towards decision making, on the basis of selected air quality information for Athens, Greece. Conclusions are drawn concerning the performance of algorithms, and for the research to be conducted in the future.

1. Introduction

Urban air quality information originates either from observations or from mathematical tools-models and estimations. While the former correspond to the current status of air quality, and may be directly interpreted in terms of human health risk and eco-system degradation potential or effect, the latter provide forecasting capabilities in advance, thus offering decision makers with the opportunity to take preventive measures that would "smooth" or alter the results of a forecasted "episode" or even "crisis". Earlier research work has dealt with using knowledge discovery techniques mainly for air quality associated incident forecasting. Several models have been built for predicting incidents that may occur in the near future. For instance, conventional statistical regression models (Bordignon et. al., 2002, Huang and Smith, 1999, Kim and Guldmann, 2001) and time-series analysis have been applied to predict ozone levels (Chen et al., 1998). Neural networks have been used for short-term ozone prediction (Ruiz-Suarez et al., 1995, Yi and Prybutok, 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 the present paper several data mining algorithms are applied, for analyzing air quality information and for forecasting the maximum ozone concentration levels on a daily basis in a dense urban area.

---

[1] MSc Programme on Advanced Systems on Computers and Telecommunications, Dept. of Electrical and Computer Engineering, Aristotle University, 541 24 Thessaloniki, Greece

[2] IDSIA - Istituto Dalle Molle di Studi sull'Intelligenza Artificiale Galleria 2, CH-6928 Manno, Switzerland

[3] Dept. of Electrical and Computer Engineering, Aristotle University, 541 24 Thessaloniki, Greece, e-mail: mitkas@eng.auth.gr

[4] Dept. of Mechanical Engineering, Aristotle University, Box 484, 54124 Thessaloniki, Greece, e-mail: kkara@eng.auth.gr

2. Data mining air quality data

2.1 The Athens air quality data set

Legislative acts, as the US Clean Air Act and the 96/62/EU framework directive for urban air management (and the accompanying daughter directives) have delimited certain thresholds for characterizing the quality for the atmospheric environment. In this background, qualitative information (like air quality indicators) is identified as being of great importance for an operational Air Quality Management System. On this basis, the current paper presents first results concerning the usage of data mining techniques for the quantitative and qualitative analysis of air quality data, and the prediction (forecasting) of pollution levels, using maximum hourly ozone concentration of a day as the parameter of investigation. For this purpose, measurement data, coming from the monitoring network operated by the Ministry of Environmental Physical Planning and Public Works were taken under consideration. The available time series of hourly values covered the period from 1/1/1999 to 1/7/2002, for the stations indicated in Figure 1, and included a numer of variables (air pollutants plus basic meteorological parameters) per station.

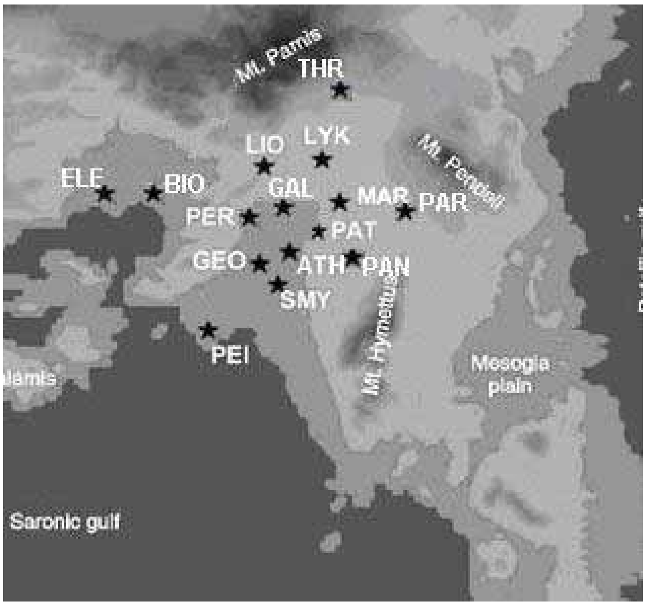

Figure 1
Locations and short names of the Athens air quality monitoring station network.

2.2 Data mining the Athens air quality data set

For all stations, the 69% of the values were used for training, and 31% for forecasting validation. Aiming at investigating the operational performance of state-of-the-art data mining methods and tools towards ozone forecasting, a number of mining methods were employed. Contrary to the conventional statistical approaches, data mining algorithms used in this paper utilize other criteria/functions, as the information gain and entropy, distance-metrics, or memory-based learning, for encapsulating data-driven knowledge and ultimately drawing conclusions. Thus, different types of algorithms for classification were applied.

Among them, instance-based learners are used, as iBK, Kstar, Nnge (Nearest Neighbor With Generalization), rule-based classifiers, as Conjuctive Rules, OneR, decision trees (C4.5 or J48 algorithm, Decision-stump), along with Bayesian Classifiers (BayesNet , NaiveBayes), Neural networks (Multilayer Perceptron) and Fuzzy-Lattice Reasoning (FLR). All these algorithms are implemented in WEKA (The Waikato Environment for Knowledge Analysis, WEKA 2004). The WEKA platform (Witten and Frank, 1999) was used (notably this is an open source software) for the data mining experiments described below.

3. Results and discussion

In order to perform the experiments concerning the forecasting capabilities of the selected algorithms, two sets of ozone limit values were applied: the one resulting from the EU experience and practice, following the relevant legislation, and the other resulting from the detailed analysis and classification of the data (Table 1).

Ozone concentration (*μg/m3*)	**Clarification**
0-44	very low
45-89	low
90-134	medium
135-179	high
180-239	information threshold
>240	alert threshold

Ozone concentration (*μg/m3*)	**Clarification**
0-29	very low
30-59	Low
60-99	low – medium
100-139	Medium
140-179	High
180-239	information threshold
>240	alert threshold

Table 1
Ozone concentration categorisation. EU oriented (left, EU), and clustering oriented (right, AUTH)

The main goal of the experiments performed was to investigate the effectiveness of the algorithms in ozone forecasting for the next 8, 24, 48 and 72 hours at each station. For this purpose, a number of success criteria were used, yet only the total weighted percentage of successful prediction is presented hereafter, for reasons of brevity. For D_i the number of available values (samples) for station I and Pi the percentage of successful prediction for station *i*, the total weighted percentage of successful prediction (TWPSP) is calculated as follows:

$$TWPSP = \frac{\sum_{i=1}^{15} D_i \times P_i}{\sum_{i=1}^{15} D_i}$$

The percentages of successful predictions resulting from the analysis are very satisfactory; a representative set of values is presented below for one of the most successful in performance algorithms (J48) (Table 2).

8EU	**24EU**	**48EU**	**72EU**
80%	86,5%	95,5%	96,6%
8AUTH	**24AUTH**	**48AUTH**	**72AUTH**
75,4%	85,4%	86,8%	89,3%

Table 2
Total weighted percentage of successful prediction of the J48 algorithm for various time frames of forecast and on the basis of two different ozone value categorisation schemes.

Overall, the best performance when it comes to successful forecasting derives from the application of the J48 algorithm, followed by Nnge and the MLP Neural Netrorks. The rest of the investigated algorithms perform within the same value spectrum.

5. Conclusions

A number of data mining algorithms have been applied for the construction of forecasting models concerning maximum per day hourly ozone concentration values, for a total of 15 monitoring sites in Athens, Greece. Successful forecasts are up to 95 %, demonstrating a good performance that should be considered for air quality forecasting modules applied at an operational basis. Future work may include the investigation of interrelations between monitoring sites and between pollutants and meteorological parameters, thus aiming to reveal "hidden" interconnection, periodicities and other qualitative and quantitative characteristics within the environmental area of interest.

Bibliography

Bordignon S., Gaetan C. and Lisi F. (2002). Nonlinear models for ground-level ozone forecasting. Statistical Methods and Applications, Vol. 11, p. 227-246

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

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

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

Ruiz-Suarez J. C., O. A. Mayora-Ibara, J. Torres-Jimenez, and L. G. Ruiz-Suarez (1995). Shortterm ozone forecasting by artificial neural networks. Advances in Engineering Software 23 p. 143-149

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

Lekkas G. P., N. M. Avouris, and L.G. Viras (1994). Case-based reasoning in environmental monitoring applications. Applied Artificial Intelligence, 8, p. 359-376.

Kalapanidas E. and N. Avouris (2001). Short-term air quality prediction using a case based classifer. Environmental Modelling and Software, 16(3), p. 263-272. WEKA,

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

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