EnviroInfo 2008 (Lüneburg)
Environmental Informatics and Industrial Ecology
Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

ICT Technologies and Somputational Intelligence Methods for the Creation of an Early Warning Air Pollution Information System

Kostas D. Karatzas[1], Anastasios Bassoukos[1], Dimitris Voukantsis[1], Fani Tzima[2], Kostas Nikolaou[3] and Stavros Karathanasis[4]

Abstract

Contemporary air quality management calls for effective, and in advance, AQ information dissemination. Such dissemination requires for communication that should not be based solely on written or oral language forms, but should make use of graphical, symbolical and multimedia language communication schemes, via available communication channels. Previous experiences and published research results indicate that the content of environmental information systems should include both real time information and forecasting for key parameters of interest, like the maximum concentration values of air pollutants. The latter are difficult to achieve, as air quality forecasting requires both domain expertise and modelling skills for the complicated phenomenon of atmospheric pollution. One of the ways to address this need and to extract useful knowledge for better forecasting and understanding of air pollution problems, is the application of Computational Intelligence (CI) methods and tools. The present paper discusses the creation of an environmental information portal for the dissemination of air quality information and warnings, for the city of Thessaloniki, Greece. The system is developed with the aid of state-of-the art, web-based technologies, including modular, on the fly software integration to operating applications, and implements CI for the forecasting of parameters of interest. In addition, observation data are made accessible to the public via an internet-based, graphics environment that deploys open source geographic information services.

1. Introduction

Early warning information services are a basic constituent for quality of life information services, and as such are inevitably associated with the atmospheric environment, the media that surrounds us throughout our life. Such services attempt to integrate the need for improved well being on a personal level with the understanding of environmental pressures and their consequences, especially at the urban scale. They also provide with valuable information concerning the way that the pattern of our everyday life is associated with exposure to, and consequences of, environmental pressures. It is becoming more and more clear that such pressures have different spatial scales (ranging from a neighbourhood to a regional problem), and multiple temporal scales (from the seconds of street canyon photochemistry to the hours of duration of a pollen episode, moving towards the days of duration of an ozone episode). The multiplicity of time and space related scales of environmental pressures calls for information services that are capable of addressing them. The present paper discusses the development of an air quality, early warning, environmental information system, that applies CI methods for the forecasting of parameters of interest, and Information & Telecommunication Technologies (ICT) for the dissemination of the environmental information. The work

[1] Informatics Applications and Systems Group, Dept. of Mechanical Engineering, Aristotle University, Box 483, 54124 Thessaloniki, Greece; email: kkara@eng.auth.gr, Internet: http://isag.meng.auth.gr
[2] Dept. of Electrical and Comp. Engineering, Aristotle University, GR-54124, Thessaloniki, Greece; email: fani@olympus.ee.auth.gr
[3] Organization for the Master Plan and Environmental Protection of Thessaloniki, Greece; e-mail: kinikola@hol.gr
[4] Directorate of Environment and Land Planning, Region of Central Macedonia, Thessaloniki, Greece; email: stkarath@rcm.gr

reported here takes into account previous systems like the European Environment Agency's Ozoneweb system (Endregard et al., 2007). Use is also made of the results of a special session organized by the first author in the frame of the ISESS 2007 conference (Karatzas, 2007), by the work conducted in the frame of another Environfo2007 paper (Trausan-Matu et al.,2007), and by a workshop organised in the frame of COST action ES0602 (http://www.chemicalweather.eu/3rdMeeting/Index). More information on urban environmental information perception and communication may be found in Karatzas and Lee, 2008.

2. Operational air quality forecasting

The core of an Air Quality forecasting system consists of a model or models, which are able to predict pollutant concentrations, in the same manner that meteorological models predict the weather. There is a great variety of models used for Air Quality forecasting purposes. These models can be grouped into the following categories

- **Human Forecasts**. In this case, data monitored in several measuring stations are collected and evaluated. A forecast is made on the basis on human experience, usually for next day's Air Quality levels. This type of forecasting cannot be considered accurate and reliable, and thus it is used only complementary to other types of forecasts.
- **Simple Statistical Approaches**, such as linear trends, moving average, exponential smoothing, ARIMA, etc. are used in many cases in order to predict Air Quality levels. Their performance varies and is influenced by the fact that the interrelations between Air Quality data are too complex for the simple statistical approaches.
- **Deterministic Models**, which can be further distinguished into *simple* models and *3-D* models. The first category includes all the approaches based on a steady-state solution of the dispersion equations, while the second category includes all Lagrangian and Eulerian models. Extensive evaluation and improvement of these types of models has been considered within the FUMAPEX project (Baklanov et al, 2007). The latest-generation steady-state models (Helsinki and Bologna) and Eulerian Chemical Transport Models (Oslo, Turin, London, and Castellon/Valencia) have been implemented into the respective air quality forecasting system, while Lagrangian models were used for emergency preparedness systems (Copenhagen). Furthermore, 3D models are currently integrated in Air Quality forecasting systems in operation, such as the UK Air Pollution Forecasting[1], PREV' AIR[2], AirQUIS[3] and others.
- **Computational Intelligence Methods**, such as Neural Networks, Classifications and Regression Trees, Self Organizing Maps, Support Vector Machines, etc. are advanced tools for knowledge discovery and forecasting parameters of interest. CI methods can be used for multiple tasks, such as classification, numerical prediction, clustering etc, while the main advantage of these methods is the accuracy combined with computational efficiency. The performance of CI methods is similar or in some cases better compared with that of deterministic models (Kukkonen et al, 2003), thus CI methods are an appropriate method to be applied for the development for operational forecasting.

3. Computational intelligence methods for AQ forecasting

Environmental data are very complex to model due to underlying interrelations between numerous variables of different type. However, as standard statistical techniques may possibly fail to adequately model

[1] http://www.airquality.co.uk/archive/uk_forecasting/apfuk_home.php?zone_id=8
[2] http://prevair.ineris.fr/fr/index.php
[3] http://www.nilu.no/aqm/airquis/

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

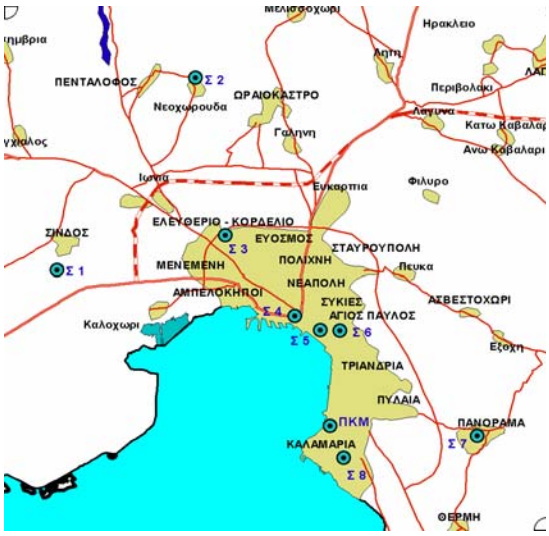
complex, non-linear phenomena and chemical procedures, the application of Computational Intelligence (CI) Methods for forecasting of a wide range of pollutants and their concentrations at various time scales, perform usually well. CI techniques such as Artificial Neural Networks (ANNs), Classification and Regression Trees (CART) and Support Vector Machines have been applied for forecasting of photochemical and particulate matter pollution in the metropolitan area of Thessaloniki (Silni et al, 2006; Karatzas & Kaltsatos, 2007; F. Tzima et al, 2007) and Athens (Grivas & Chaloulakou, 2006; Chaloulakou et al, 2003; Athanasiadis et al, 2006). The results of these studies indicate that CI methods perform better compared to statistical methods, and can be potentially very accurate in forecasting parameters of interest, depending on the quantity and quality of the data. These findings, combined with the computational efficiency of CI methods, suggest that the latter methods can be an excellent tool for the creation of operational air quality forecasting modules, which may effectively support operational air quality management on a day-to-day basis.

4. Construction of operational prediction models

Data mining (DM) has been defined as the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al. 1992). It employs various CI techniques (supervised or unsupervised learning algorithms), in order to automatically search large volumes of data and derive patterns that can be used for either predictive (classification/regression) or descriptive tasks (association rule mining, clustering, etc.). In the context of our current work, DM is used for classification that can be formally defined as "....the task of learning a target function f that maps each attribute set x to one of the predefined class labels y" (Tan et al. 2005).

In the following, we provide with a short overview of the computational experimental strategy employed in order to construct operational prediction models for the city of Thessaloniki, which is the second largest city of Greece (more than one million inhabitants) and one of the largest urban agglomerations in the Balkans. Its complex coastal formation, in combination to the near-by mountainous areas, forms a very complex land use and orography pattern that favours local circulation systems. Thus, the formation and transport of pollutants are heavily influenced by the local meteorological and topographic characteristics, which is the case in many coastal urban areas around the world. In this work, use was made of datasets of meteorological and air pollutants hourly measurements (more than 200,000 records), supplied by the Directorate of Environment and Land Planning of the Region of Central Macedonia (RCM), Greece. The map of the monitoring network is presented in Figure 1. After removing records with missing values, as well as obviously erroneous measurements, a simple method of linear interpolation was applied, for estimating missing values when the "time gap" was less than 48 hours. Moreover, numerical values for all pollutants (including the class variable) were transformed to nominal values.

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2



Station number	*Station name*
Σ1	TEI Sindos
Σ2	Neohorouda
Σ3	Kordelio
Σ4	Dimikratioas Sg.
Σ5	Ag. Sofias Sq.
Σ6	AUTh
Σ7	Panorama
Σ8	Kalamaria

Fig. 1: The Thessaloniki air quality monitoring network of the RCM.

All computational experiments were conducted with the aid of the Waikato Environment for Knowledge Analysis (WEKA) (Witten and Frank, 2005) and involved algorithms falling in the following categories: (i) *Decision Tree Classifiers* (Decision Stump, Logistic Model Trees, Naive Bayes Trees, Random Forest, Random Tree, REP Tree and J48 – WEKA's C4.5 implementation); (ii) *Neural Networks* (Multi-Layer Perceptron and RBF Network); (iii) *Rule-based Classifiers* (JRip, OneR, Conjunctive Rule, Decision Table, Ridor, PART and NNge); (iv) *Bayesian Classifiers* (Bayes Network, Naive Bayes and Naive Bayes Updateable); (v) *Instance-based classifiers* (IBk, IB1, KStar); (vi) *Support Vector Machines* (sequential minimal optimization algorithm – SMO); and (vii) *Logistic Regression* (Logistic, Simple Logistic).
Our experimental strategy is presented in the next chapter and involved three stages:

- Data pre-processing and formulation of appropriate datasets for each prediction task
- Evaluation of alternative classification algorithms per station and per prediction task
- Construction of cost-sensitive operational prediction models for the best performing algorithms of stage 2 and evaluation of these models (i) using 10-fold cross validation and (ii) on the complete training set.

5. Computational experimental strategy

We have used eight datasets of meteorological and air-pollutants measurements, combined with seasonal information for estimating (i) *daily maximum* NO_2, (ii) *maximum* O_3 and (ii) *mean and maximum* PM_{10} *concentration levels*. The datasets were supplied by the Directorate of Environment and Land Planning of the Region of Central Macedonia, Greece, and come from monitoring stations in the metropolitan area of Thessaloniki that record several meteorological attributes and air-pollutant values on an hourly basis.

After formulating the appropriate datasets for each prediction task – selecting the subset of parameters used for forecasting each of the target pollutants and calculating the corresponding values from the hourly concentrations present in the original datasets – we removed records with missing values in the class attribute (the maximum or mean concentration value to be predicted) and transformed the class attributes to nominal values, in accordance to relevant technical guidelines defining the corresponding scales.

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

Then, we proceeded by evaluating almost all available (and applicable) algorithms within the WEKA environment, using limited computational resources and within a reasonable time frame. For each of the stations and for each of the target pollutants, several experiments were conducted, each evaluating a specific algorithm using 10-fold cross validation and 5 repetitions (thus providing a statistically significant performance measure as the mean over 50 evaluations per station-algorithm pair). The final product of this procedure was a list of the 5 best performing algorithms per station and target pollutant.

Having selected the top performing algorithms, we observed that none of them managed to predict exceedances (HIGH and VERY HIGH class values) accurately: due to the scarcity of records representing above-threshold concentrations and the complexity of the domain we are trying to model, only one tenth of constructed models managed to achieve a prediction accuracy greater than 50% in the case of exceedances. To circumvent this shortcoming, we decided to employ cost-sensitive model building for our final operational models. In this direction, a cost-matrix was defined, based on the knowledge domain of the authors that mirrors the relative cost each of the model's misclassification has:

	Predicted Class			
Actual Class	**LOW**	**MEDIUM**	**HIGH**	**VERY_HIGH**
LOW	0	5	20	35
MEDIUM	5	0	15	30
HIGH	40	35	0	55
VERY_HIGH	105	100	105	0

- 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
- 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.

Cost-sensitive model building proved to be an effective technique to "guide" the algorithms towards the intended outcome: all models built using the above cost-matrix managed to outperform the corresponding non-cost-sensitive ones (comparison per algorithm-station-target pollutant triplet) achieving prediction accuracies up to 88.5% for exceedances. The algorithms that were proven to be more effective were mostly in the category of decision trees, and more specifically J48 (WEKA's C4.5 implementation) and Logistic Model Trees. More details are available in Tzima et. al., 2007.

6. The early warning information system

Air quality information systems have already been addressed the era of the 4th FP in EU. The first communication channels investigated were the ones supported by internet technologies. In the 5th FP a number of IST related projects addressed air quality management, information, and systems. The reference project was APNEE and its take-up measure APNEE-TU[1] that addressed, for the first time, the needs of the citizens for personalised information services for the quality of the environment they live in, and developed an umbrella of pull and push services that can be used for providing AQ information to the public. APNEE and APNEE TU (2000-2004) provided with a holistic approach to AQ information management and dissemination. A number of projects followed, while in parallel operational systems started to emerge. Some examples include the following systems:

[1] http://www.apnee.org

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

- Luftkvalitet[1]. The official Norwegian AQI site, developed and supported by the Norwegian Institute for Air research in the frame of their air quality management system AIRQUIS[2] and its information component AirOnline, that provides web based, mobile phone and street panel information dissemination
- AirNow[3], the official air quality information system of the Environmental Protection Agency, USA. This system received input from more that 3000 monitoring stations, and provides information via the internet and with emails.
- AirALERT[4], which is an SMS based air quality information service for the Sussex area, U.K., especially focusing on asthmatic people, school children and elderly.

The Thessaloniki air quality information and early warning system (AIRTHESS), was designed and developed taking into account the state of the art in ICT and the way that AQ information should be disseminated and presented, either on a daily information basis, or on the basis of alerts generated by incident forecasts. AIRTHESS makes use of Google Maps[5] for the geographic presentation of information and Adobe Flash for the graphical presentation of air pollution time series, merged in a responsive rich web application using the Google Web Toolkit (GWT)[6] to perform dynamic actions. In the backend, AIRTHESS is implemented using a stack of open source libraries and frameworks, mainly based on the Eclipse Equinox implementation of the OSGi[7] Service Platform, allowing fine-grained reuse of existing code and extension of behaviour with minimal overhead. Further, an in-house web application framework based on Apache Velocity[8] for rendering and Apache Torque[9] for the database access. AIRTHESS uses the OSGi Event Service in publish-subscribe mode to handle dataflow requirements for new measurements and notifications. When new measurements arrive, an Event is generated describing their metadata (such as their station, the measurement series that have been updated, the first and last moment of the new data, etc) and asynchronously posted to the Event Service. Any OSGi services that have registered as handlers whose filters match the new event are then notified; the modeling subsystem is one such service, handling execution of prediction models. Any forecasts that are calculated emit their own events; the notification service is subscribed to these particular events and checks if any users should be notified of these new predictions. Being based on OSGi and the publish-subscribe pattern allows the various components to be very decoupled from each other; further, services can dynamically subscribe to the generated events without causing any downtime. The warning are being issues via e-mail and SMS to the (freely) subscribed users. On the software development side, we have also investigated Apache Camel[10] as a JVM-bridging transport for the OSGi Event Service, which should allow us to host the prediction models in a second JVM with failover and perhaps better performance.

7. Conclusions

The dissemination of environmental information on a regular basis but also on the basis of event-based alarms is of major importance for public environmental administrations and citizens alike. As the quality

[1] http://www.luftkvalitet.info/
[2] http://www.nilu.no/airquis/
[3] http://airnow.gov/
[4] http://www.sussex-air.net/airalert.html
[5] http://code.google.com/apis/maps/
[6] http://code.google.com/webtoolkit/
[7] http://www.osgi.org/
[8] http://velocity.apache.org/
[9] http://db.apache.org/torque/
[10] http://activemq.apache.org/camel/

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

of the everyday life and the protection of public health and safety are in the top of the agenda for every modern governance, the introduction of smart, location oriented, flexible and adaptable information services is becoming more and more a necessity. In addition developments in this area in both Europe and the USA, have shown that the citizens are interested in receiving such information and in having access to such services and data. On this basis, an air quality, early warning information system was developed for the city of Thessaloniki, Greece, under the name AIRTHESS (www.airthess.gr). The system incorporated computational intelligence methods for the forecast of pollution levels, and state of the art technologies for the presentation and disseminating of both every day information and warnings. The system is easily adaptable to local needs (in both ICT and data infrastructure), and was specially designed so that it may be applied for other environmental and public administration domains. The first operational tests have verified the expectations of the design team, while its everyday operation is expected to result in the improvement of the public's knowledge on the quality of the environment they live in.

8. Acknowledgements

The authors acknowledge the project AIRTHESS, supported by the Organization for the Master Plan and Environmental Protection of Thessaloniki, and the Region of Central Macedonia, Thessaloniki, Greece. The work of this paper is also related to COST action ES0602 (www.chemicalweather.eu), and COST action C21 (www.towntology.net).

References

Athanasiadis I., Karatzas K. and Mitkas P. (2006), Classinfication Techniques for Air Quality Forecasting, Proceedings of the BESAI 2006 Workshop on Binding Environmental Sciences and Artificial Intelligence (Oprea M., Sances-Marre M. and Wotawa F., eds., part of the 17th European Conference on Artificial Intelligence- ECAI 2006), pp. 4-1 to 4-7.

Baklanov A., O. Hanninen, L. H. Slørdal, J. Kukkonen, N. Bjergene, B. Fay, S. Finardi, S. C. Hoe, M. Jantunen, A. Karppinen, A. Rasmussen, A. Skouloudis, R. S. Sokhi, J. H. Sørensen, and V. Ødegaard (2007), Integrated systems for forecasting urban meteorology, air pollution and population exposure, *Atmos. Chem. Phys.*, **7**, 855–874,

Chaloulakou A., Saisana M and Spyrellis N. (2003), Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens, *The Science of the Total Environment* **313**, 1–13

Endregard G., Karatzas K., SkaanesB.I., Fløisand I. and Larssen S. (2007), EEA air quality web dissemination solution - recommendations for further development, ETC/ACC Technical Paper 2006/9. Report prepared for the European Environment Agency, http://air-climate.eionet.europa.eu/docs/ETCACC_TechnPaper_2006_9_AQ_web_dessim.pdf

Frawley W., Piatetsky-Shapiro G., and C. Matheus (1992) Knowledge discovery in databases: An overview, *AI Magazine* **13**, 57-70.

Grivas G. and Chaloulakou A. (2006), Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece, *Atmospheric Environment* **40,** 1216–1229

Karatzas K. (2007), Session on Environmental Engineering Education and Presentation of Environmental Information to Non Scientists, ISESS2007, http://www.isess.org/

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

Karatzas K. and Lee J. (2008) Developments in urban environmental information perception and communication, iEMSs 2008: International Congress on Environmental Modelling and Software, Integrat-

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2

ing Sciences and Information Technology for Environmental Assessment and Decision Making. Proceedings (in press), M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)

Kukkonen J., L. Partanen, A. Karppinen, J. Ruuskanen, H. Junninen, M. Kolehmainen, H. Niska, S. Dorling, T. Chatterton, R. Foxall and G. Cawley (2003). Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment* **37** (32), 4539-4550.

Slini T., Kaprara A., Karatzas K. and Moussiopoulos N. (2006), PM10 forecasting for Thessaloniki, Greece, *Environmental Modelling & Software* **21,** 559–565

Tan P. N., M. Steinbach, and V. Kumar (2005) Introduction to Data Mining (First Edition), Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Trausan-Matu S., Karatzas K. and Chiru C. (2007), Environmental information perception, analysis and communication with the aid of natural language processing, Proceedings of the 21st International Conference on Informatics for Environmental Protection - EnviroInfo2007, (Hryniewicz O., Studziński J. and Romaniuk M., eds.), Vol. 1., pp. 299-306, Shaker Verlag, Aachen, 2007, ISBN 978-3-8322-6397-3 (conference date and location: Warsaw, Poland, Sept. 12-14, 2007)

Tzima F., Karatzas K Mitkas P and Karathanasis S. (2007), *Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece*, Proceedings of the 20th International Joint Conference on Neural Networks, Page(s):2752 – 2757. Organized by the IEEE Computational Intelligence Society and by the International Neural Network Society, Orlando, Florida, August 2007

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

Copyright © Shaker Verlag, Aachen 2008. ISBN: 978-3-8322-7313-2