

# Knowledge Based Environmental Data Validation

Stefania Bandini<sup>a</sup>, Davide Bogni<sup>b</sup>, and Sara Manzoni<sup>a</sup>

<sup>a</sup>Department of Computer Science, Systems and Communication, University of Milano–Bicocca, Via Bicocca degli Arcimboldi, 8, - 20126 Milan (ITALY), e.mail:manzoni@disco.unimib.it

<sup>b</sup>Project Automation S.p.A., Viale Elvezia, 42 - 20052 Monza (MI)

**Abstract:** This work describes EcoExpert, a knowledge–based module supporting data validation and integrated into a system for environmental data management. The paper describes the integration of knowledge–based system technology and data monitoring networks for environment management purposes by means of influence graphs. EcoExpert has been designed to communicate directly with the air quality monitoring network managed by a Pilot Operating Unit. It is considered an example for achieving an integrated environment management system using knowledge–based technology and qualitative reasoning techniques to solve problems involving data validation, intelligent alarms filtering, network diagnosis, network configuration and maintenance scheduling.

**Keywords:** knowledge–based monitoring, environmental data management and validation

## 1 INTRODUCTION

The main aim of this work is to describe a knowledge–based system (EcoExpert) supporting data validation. EcoExpert is integrated into a general computational framework dedicated to environmental data management. In particular, EcoExpert has been designed to support professional staff dedicated to validate data collected by a network of remote pollution data monitoring tools. Data validation consists in two main steps. First, data collected by the remote sensors are processed to identify possible anomalies, representing polluted air situations but also faults in the sensing tools. Then, each identified set of suspect data is correlated to the set of its possible causes, and those not effectively correlated to any cause are invalidated.

Given a set of anomalous data the EcoExpert module supports the data validation process both in data processing and in finding out chemical, physical or human conditions that could justify the detected anomaly. Two types of representations have been adopted for the EcoExpert knowledge base: *influence diagrams* and the *heuristic knowledge model*. Influence diagrams support the numerical–statistical data analysis, describing the relationship among the concentrations of different types of pollutants and between the concentration of a pollutant and the state of remote sensing tools. The heuristic

classification model supports the correlation of each anomaly identified in data about pollutants to their possible causes, taking into account the monitoring log of the sensing tools, data about atmospheric conditions and other correlated pollutants, and the topology of the network of sensing tools. This rule–based representation models expert knowledge and it has allowed to support EcoExpert users (possibly not expert in analysis techniques) in the data validation process. Every day the EcoExpert system supports the professional staff of various Italian provinces and regions in validating data collected the previous day by a network of sensing tools and in justifying eventually identified anomalies.

In the last few years, many different information technology applications have focused on environmental issues. Operators in the field have shown a growing interest in solutions designed to assess, prevent, and reduce the impact on the environment of human activities, or to monitor environmental conditions (Gerelli [1990], Lekkas et al. [1994]). Given the complexity of the processes to be analyzed and the number of factors involved, decision–makers can gain considerable advantage from using automatic tools currently available to support some of the work deriving from the intricate panorama of environmental management (Guarisio and Werthner [1989]). Moreover, information technology provides a testing ground for simulations that is free of all the time/cost restrictions and risks that real ex-

periments would involve.

The basic problems facing environmental management today include:

- acquisition of data that are relevant to the problem (dynamic and/or static);
- data organization and storing in order to make it easily readable;
- reasoning over it with the aid of knowledge-based techniques or predefined models in order to recognize dangerous situations, identify the causes of specific environmental degradation, predict the impact of human activities on the environment, and planning of urban and industrial development.

We can picture the technology for environmental control as a pyramid, where the base is made of telecommunication technologies, with monitoring networks collecting data about weather and pollution conditions. Data acquired in this way are transmitted to supervision centers, and, if needed to operators and users.

Database management systems and geographical information systems (GIS) technologies form the next level up. In this level, data required for environment management are organized and made easily readable by other system components and by system users; such data include:

- meteorological data (rainfall, solar radiation, air temperature, relative humidity, wind speed and direction, hydrometric level, and so on);
- pollution data (e.g., concentration of  $SO_2$ ,  $NO_2$ ,  $CO$ , Suspended Particles, hydrocarbons);
- static data (e.g., residential and industrial buildings, census data, regulations, land registers);
- information system management and maintenance data.

The third level of the pyramid is concerned with the problems that arise once the data provided by the lower levels have been processed. The work at this level is the most critical for environmental control organizations, since it requires the competencies of a highly qualified and experienced professional staff. The main activities of this staff can be summarized in: validating incoming data from the network, making network diagnoses, planning maintenance

work on monitoring and network instrumentation, summarizing, elaborating and presenting the data in a differentiated way to the users (local administrators, engineers, operators, the public), making environmental impact assessments, studying correlations among different pollutant factors, proposing regulations, designing and constantly updating a network configuration that best represents the realities to be examined, selecting the most appropriate models (e.g., economic, physical, ecological, meteorological).

People performing this work need great experience. This is the reason why these issues are suitable for the application of expert system technology, which enables knowledge and qualitative reasoning to be treated as data. In this way, a greater degree of abstraction that is essential for dealing with this kind of problem can be obtained. As a matter of fact, the role of expert systems as integration tools in heterogeneous automation environments is increasing (Terplan [1986]).

In this paper we present an experience in this field, concerning the automation of data validation, that is one of the most critical jobs immediately deriving from data collection. Data validation concerns checking whether data coming from a network are correct and reliable. The work was done for Pilot Operating Units (UOP) which run air quality data monitoring networks in heavily polluted urban areas, with the development of an expert system (Eco-Expert), designed to validate data concerning air quality.

## 2 THE DATA VALIDATION PROBLEM

All the work concerning environmental management is first of all based on data concerning pollution levels and atmospheric conditions in a given area. These data are automatically collected by a monitoring system and sent to the UOP supervisory center where they undergo validation checks. Validating means checking whether the data recorded by remote sensors in the field and sent to the supervisory system are not affected by errors. In particular, data collected by the network monitoring system are validated every day, and a report providing a summary of the condition of air pollution of a given area is sent to many different public and private institutions.

The data validation of daily data, collected in twenty-four hours, can be divided in three distinct stages:

1. Numerical-statistical processing of the recorded data according to conformity cri-

teria, such as: maximum hour test, high difference test, spike test, high consecutive four hours test, and flat series test (the first four points are suggested by the Environmental Protection Agency (Agency [1978])). The aim of this stage is to single out any odd-looking data needing further investigation.

2. Trying to find out the cause of the numerical difference in the suspicious data according to various possible causes (type, location and conditions of the instruments, atmospheric conditions, and so on) before definitely declaring the data not valid. The factors, that are considered in order to explain the odd values obtained for the suspicious data, vary significantly according to the type of pollutant in question (e.g., monitoring site, season, time of day, type of monitoring instrument, maintenance history, meteorological and human factors).
3. If the above operation has not managed to find a cause for the odd-looking data, or if it has identified a malfunctioning of instrumentation as the cause, the data are effectively invalidated.

### 3 THE ROLE OF THE EXPERT SYSTEM

It is obviously vital for the UOP and all the operators involved in monitoring and assessing pollution to base their decisions on reliable data. It is therefore essential for the methods and techniques used to collect and validate data to be as simple as possible. On the other hand, such methods must guarantee absolute reliability, so that the data published and subsequently used for study, modelling and forecasting purposes are undoubtedly correct. At the moment, there is no effective system allowing to diagnose any possible problem on the monitoring instrumentation or on the network quickly, remotely and automatically. Thus, the reliability of the instruments and, consequently, of the data collected can be assessed only with on-site controls. This means that the UOP people have to assess the effective validity of the data measured according to their own experience and knowledge of the measuring instrumentation.

The critical nature of the controls made by the UOP is self-explained if one considers the organizations, public authorities and public opinion that base their decisions on environmental data. An expert system dedicated to the support of this community of operators has to preserve the intrinsic nature of the data

needed by operators, and at the same time must integrate different tools providing the data. We oriented our choice towards an integrated solution, in order to satisfy the requirements and the specific functional requirements. Because of the versatility of the techniques available nowadays in the framework of knowledge-based technologies, we focused our attention on rule-based technologies (for capturing heuristic knowledge) and on a qualitative reasoning approach. The latter choice has been made in order to describe at the conceptual level the knowledge involved, and in order to perform numerical computations to obtain data by inference whenever data are not directly available.

#### 3.1 The EcoSystem

EcoExpert belongs to the EcoSystem, an integrated system performing supervision, management and control of an air quality monitoring network. As shown in Figure 1, such a the EcoSystem consists of a set of applications. This applications can be classified according to the task they perform. First, a set of devices are devoted to data acquisition, transferring and storage into a centralized data base (respectively, EcoRemote and EcoManager modules). Moreover, other modules are dedicated to network configuration and management (i.e. EcoEdit), data post-computing and reporting (i.e. Analyzer and EcoNet) and data validation (i.e. EcoExpert).

The EcoSystem architecture is quite complex and consists of three main levels: central, supervising and peripheral levels. At the *peripheral level* a set of monitoring devices are dedicated to data acquisition, analysis and sending to the *central level* in order to be validated, stored and used as source for pollution reports (*supervising level*). The EcoRemote module is dedicated to the numerical-statistical analysis of data acquired by the monitoring sensors. All the suspicious-looking data are identified both with numerical analysis tools (e.g., by checking maximum and minimum, standard deviation of the values) and by comparing the data against predefined threshold values. The EcoManager module manages the environmental data base (i.e. EcoDB) that contains all the relevant static and dynamic data of the system. Static data include network configuration data and test configuration data, while the dynamic ones concerns the maintenance data and air pollution concentration data coming from the remote sensing systems. Dynamic data can be updated either at regular predefined intervals (e.g., every few hours) or by validation operations, automatically performed by the inference en-

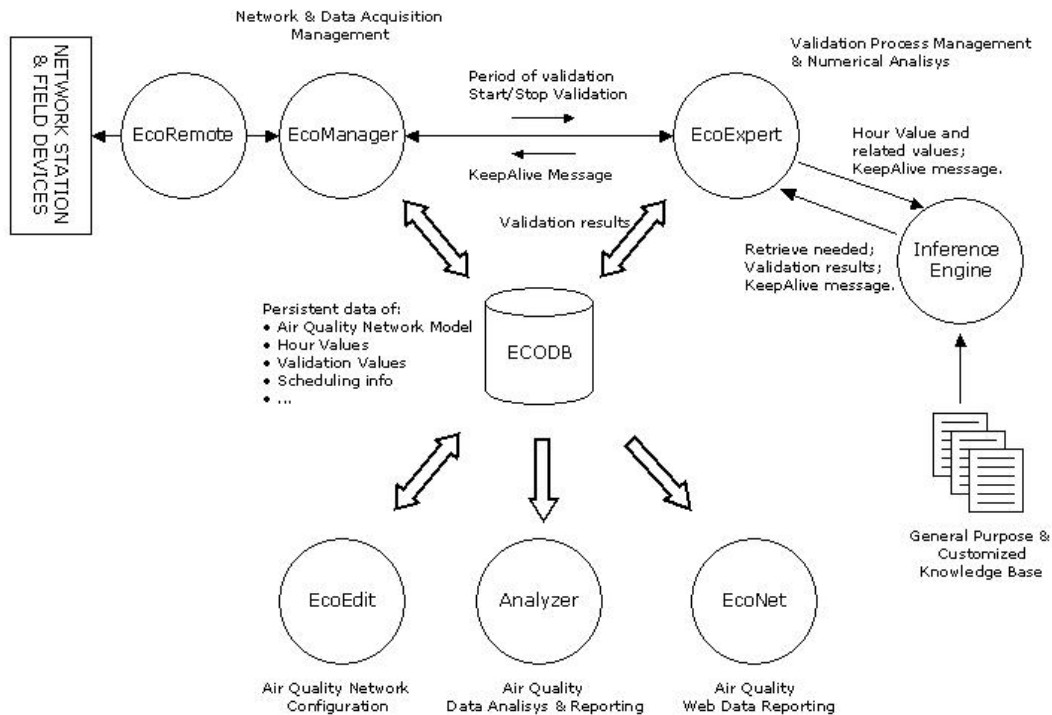


Figure 1: The EcoSystem Architecture.

gine of the EcoExpert module. The user interface of the EcoEdit module allows the management of all the operations involving final users and supports the graphical user interface. It is composed by three main subsystems: the manual analysis and validation sub-system (graphical and numerical), the test configuration subsystem, and the subsystem that reports all the activities.

### 3.2 Heuristic Analysis of Data: the EcoExpert module

A priori, all data collected by the network have to be considered valid. Data transmission and archiving systems are reliable, thus only the correct functioning of the sensors that collect and record data has to be analyzed. Unfortunately, network devices do not provide powerful auto-diagnosis capabilities, and the remote access to the functional status of the various monitor components is still limited. Thus, the main problem in the development of an automated support in this domain is to determine whether anomalies in the data are actually caused by air pollution or are the effect of the malfunctioning of the remote sensors. This is the goal of the EcoExpert module.

The software architecture of EcoExpert is characterized by a main component that manages the communication interface (i.e. socket) with the EcoManger module in order to receive a set of parameters needed for its elaboration (e.g. data to be validated and validation start time). Only after a complete set of data has been acquired and stored by the EcoManger, the main component of EcoExpert can acquire data from the EcoDB. The first elaboration phase is the numerical-statistical analysis of this data in order to identify specific anomalies. After the storing of the result of this first elaboration, a second elaboration phase performs the heuristic analysis. The aim of the heuristic analysis is to try to justify every identified anomalies. This task is performed according to correlations among pollutant concentrations acquired by different peripheral devices, concerning different pollutants or different time periods. The conceptual model adopted for the representation of the knowledge involved in the heuristic analysis about pollutant correlations will be described in the following section.

## 4 THE CONCEPTUAL MODEL

If the appearance of anomalous data cannot be explained with enough certainty by a failure of some network sensor, the idea is to, conversely, assume that the data are valid and the network instruments are working properly. Therefore, the system tries to determine whether the presence of suspicious data can be related to physical–chemical and/or human conditions.

### 4.1 Strategic Knowledge

The physical data validation process consists of an examination of all the data collected, in order to identify and single out all the elements that could be the results of errors or malfunctioning. Once one, or more, anomalies have been identified, the system first checks monitoring instruments. The reason for the anomaly, in fact, might be a failure in the instrumentation, due to maintenance problems of monitors, low monitor reliability, status of the monitor when the measurement was made. For each of these conditions one or more possible causes are selected and checked (for example, an instrument not subject to frequent maintenance, a monitor undergoing maintenance during the measurement), with a procedure that generates and test different hypotheses. If none of the above causes clearly seems to be responsible for the suspicious data, analysis process continues, by suggesting some external situations that might have given rise to the unusual variation of the value recorded. These solutions are based on the information of the experts of the field. In this case, the expert must distinguish between phenomena that could justify an increase in the pollutant and others that could make it decrease. These phenomena can be classified into accumulation and reduction processes respectively. In both cases the processes may have a physical–chemical or a human origin. The identification of what caused the anomaly in the data may eventually lead to the direct invalidation of part the data collected by that particular monitor, depending on the type of cause identified (for example an unreliable monitor), even if the latter had not been indicated as suspect.

The knowledge involved in this process has been represented by means of an influence graph (Ad-danki et al. [1991]), which represents data dependencies. The nodes of the graph define the possible states of both the concentration of pollutants and the functioning of the remote sensing tools. The nodes can be connected by two types of directed edges: *comparison edges* and *verification edges*. Comparison edges link correlated pollutant concentrations

(i.e., the homogeneous increase of NO<sub>2</sub> in urban areas implies the increase of CO). Verification edges can define both a link between a pollutant value and the state of the remote sensing tool, and between a pollutant set of values and temporized boundary conditions (e.g., time, season).

The influence graph as been developed according to both general domain knowledge (i.e. correlation between pollutants, position of monitoring devices, weather conditions, and so on) and specific and customized knowledge characterizing each monitoring device. While the first type of knowledge has been acquired from domain expert during the system design, the latter is the result of a second knowledge engineering process that has been performed after the EcoSystem installation. According to user requirements, in fact, specific features characterizing some of the devices have been included into the knowledge base of the EcoExpert system. Examples of these features concern, for instance, specific information about device location and information about reliability of network functioning and maintenance scheduling that can be customized by system users.

### 4.2 Heuristic Knowledge

The characterization of the situation identified is correlated heuristically to one or more possible classes of causes (processes) capable of explaining the anomaly in question. By a process of refinement the various hypothetical interpretations are then examined one by one, until the cause or causes at the root of the phenomenon are identified. Once the cause of the phenomenon observed has been identified, the data in question can be classified, by definition, as valid or non valid. The classical model for representing this kind of knowledge has been the heuristic classification model (Clancey [1985]). It has been implemented by a rule–based development tool and consists in a separate module which examines the suspect data (coming from the numerical–statistical analyzer) in order to justify the anomalies.

In the following, two example rules are shown. The first one represent a typical data validity rule. The aim of this type of rules is to represent conditions (e.g. correlation between anomalous pollutant concentration and location of the monitoring devices) that can explain an anomaly in acquired data. For instance, the rule below justifies an homogeneous increase of pollutant whose primary source is intensive traffic if the monitoring device is characterized by intensive traffic at the time of data acquisition.

Under this conditions the system judges as a normal case of pollutant accumulation the high increase in the pollutant concentration. Conversely the second rule is a typical non-validity rule. In this case a problem in the monitoring tool and in its maintenance is defined as the cause of an increase of pollutant concentration if it not contemporaneously present an increase of concentration of pollutants characterized by the same primary source.

<p>RULE m: A Data Validity Rule  IF  at time t a homogeneous increase of pollutants belonging to a station is present  AND  the pollutants whose increase is measured have as primary source the condition of intensive traffic  AND  the expected time is maximum traffic time  THEN  there is a accumulation process</p> <p>RULE n: A Data Non-Validity Motivation Rule  IF  at time t an homogeneous increase of pollutants is present  AND  the pollutants whose increase is measured have as primary source the condition of thermal plants  AND  homogeneous increases of correlated pollutants are not monitored  AND  the station is in a rural area  THEN  there is a problem in the monitoring tool maintenance</p>
---

## 5 CONCLUSION AND FUTURE DEVELOPMENTS

The EcoSystem is an example of an integrated system for the environment, which, alongside the typical environmental monitoring and control system functions (remote monitoring, data archiving and presentation), provides new facilities which make the system a real and valuable aid at all stages of environmental management. Moreover the EcoExpert module confirms that expert system technology is able to provide concrete solutions to problems with high intrinsic complexity, that require great experience from human experts and that lack complete theoretical structuring; such characteristics are typical of the problems surrounding environmental management, starting from the task of validating automatically monitored data.

The technology involved into the EcoSystem are the ones commonly available, and specifically:

- hardware consists of high quality Personal Computers with multitasking operating system (Windows 3.11 characterizing the original system configuration has been recently upgraded to Windows 2000 Server/Professional);
- software development, with C++ and RAD programming languages (e.g. Visual Basic), has followed the Object Oriented paradigm providing the whole system with modularity, scalability and high configurability;
- communication protocols between applications, even on various supports (serial link, local network, and so on), are all based on the TCP/IP standard;
- the Data Base Management System is based on the relational model.

Following the successful outcome of this first step, and in view of achieving the system, EcoSystem will shortly be equipped with further expert functions such as routine and extraordinary network management, automatic, daily generation of work schedule for maintenance staff, intelligent alarm tools.

## REFERENCES

- Addanki, S., R. Cremonini, and J. Scott. Graphs of models. *Artificial Intelligence*, 51, 1991.
- Agency, E. P. Screening procedures for ambient air quality data. *Technical Report*, 1978.
- Clancey, W. J. Heuristic classification. *Artificial Intelligence*, 27:289-350, 1985.
- Gerelli, E. Ascesa e declino del business ambientale. *Ed. 11 Mulino*, 1990.
- Guarisio, H. and Werthner. Environmental decision support systems. *Ellis Horwood Limited*, 1989.
- Lekkas, G. P., N. M. Avouris, and L. G. Viras. Case-based reasoning in environmental monitoring applications. *Applied Artificial Intelligence*, 8, 1994.
- Terplan, K. Expert systems for network operational control. *Proceedings of CMG-USA*, 1986.