Identification of redundant sensors in an air pollution network using cluster analysis and SOM

G. Ibarra-Berastegi[1,2], J. Sáenz[1,3], A. Ezcurra[1,3], U. Ganzedo[1,3], A. Elias[1,4], A. Barona[1,4] & A. Barinaga[1,2]
[1]*University of the Basque Country, Spain*
[2]*Fluid Mechanics & N.I. Department, Faculty of Engineering, University of the Basque Country, Spain*
[3]*Applied Physics II Department, University of the Basque Country, Spain*
[4]*Environmental & Chemical Engineering Department, University of the Basque Country, Spain*

Abstract

An air pollution network monitors – among others – the sulfur dioxide (SO_2) levels at 4 locations in Bilbao city (Spain) and surroundings. The main objective of this work was to develop a practical methodology to identify redundant sensors and evaluate the network's capability to correctly represent SO_2 fields throughout the whole area. The methodology is developed and tested at this particular location, but it is general enough to be useable at other places as well, since it is not tied neither to the particular geographical characteristics of the place nor to the phenomenology of the air quality over the area. To that purpose, the combination of two different techniques has been used: Self-Organizing Maps (SOM) and cluster analysis (CA). The results show that both techniques yield the same results, but the information obtained via SOM can be helpful not only for that purpose but also to throw light on the major mechanisms involved. This might be used in future network optimization stages. The main advantage of CA and SOM is that they provide readily interpretable results.

Keywords: sulfur dioxide, air quality network, cluster analysis, Self-Organizing Maps, spatial variability, Bilbao, fluid mechanics, applied physics, chemical engineering.

WIT Transactions on Ecology and the Environment, Vol 136, © 2010 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR100321

1 Introduction

If a network is properly maintained and periodically evaluated, its measurements can be helpful to inspire further policies and draw solid conclusions on several aspects like for example, trends.

In the case of SO_2, in the last years, a downward trend has been reported in Spain as well as in most European countries. Bilbao's network is not an exception regarding these general trends (Ibarra-Berastegi et al. [1–5]). Since 1977, throughout the whole ZONE B (Fig. 1), an air pollution network monitors the evolution of several pollutants. Emissions affecting ZONE A are scattered throughout the city, domestic heating systems being the major SO_2 sources. However, in ZONE B, additional sources originate important emissions and impacts.

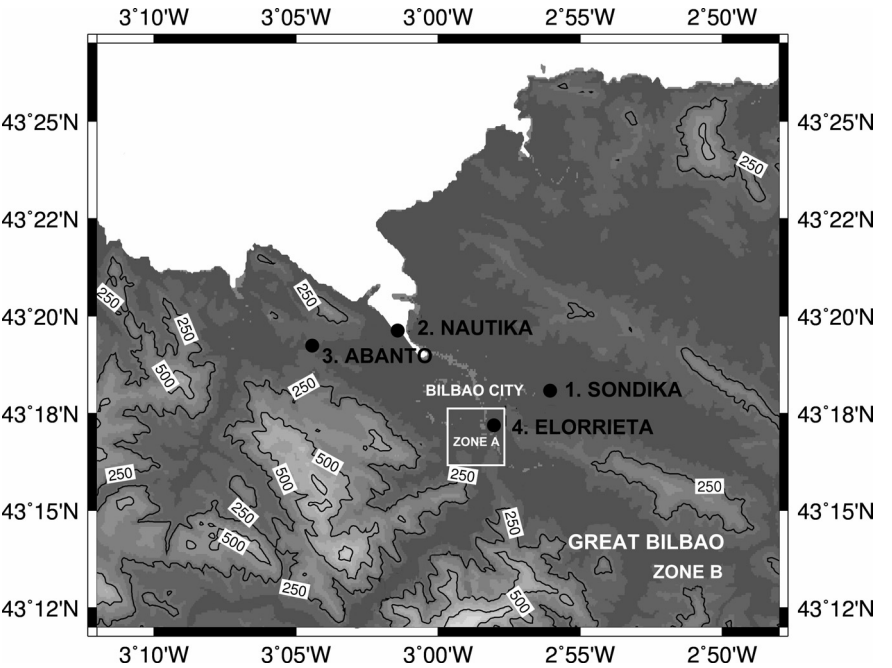

Figure 1: Area of the study.

For the case of ZONE A (Fig. 1), it has been shown (Ibarra-Berastegi et al. [5]) that the six existing sensors capture the same SO_2 regimes, that is, the six provide redundant information. The explanation for this is that the emission-dispersion-inmission mechanisms inside the area of Bilbao city can be described following the single box model where the SO_2 is almost perfectly mixed in the boundary layer above the city (Ibarra-Berastegi et al. [5]). The single box model (Zannetti [6]) is based on the mass conservation of pollutants inside a Eulerian

box. As a consequence, the six sensors of ZONE A see nearly the same (Ibarra-Berastegi et al. [5]) and in this sense, it can be stated that they are redundant.

In the last years, air quality control measures have been implemented in the area and, like in any other network, evaluation of emission abatement policies are required. This can only be done if an air quality network with an appropriate configuration is available.

The objective of this paper is to develop a practical methodology to identify redundant sensors and evaluate a network's capability to correctly follow and represent SO_2 fields throughout a whole area (ZONE B in Bilbao for this study, Fig. 1). To that end, one of this group of six sensors inside ZONE A (namely Elorrieta) city has been chosen as representative of Bilbao city and compared with 3 other sensors (Sondika, Nautica and Abanto) located in ZONE B.

If no significant spatial variability is detected among the measured SO_2 levels at these 4 locations, it would mean that the four sensors provide redundant information and that the air quality network needs further changes in its configuration.

2 Methodology

For this study, 68 monthly records from four SO_2 sensors in the Great Bilbao area (ZONE B Fig. 1) corresponding to the 1996-2001 period have been used The present methodology includes the combined use of SOM and cluster analysis.

2.1 Self-Organizing Maps

The Self-Organizing Maps (SOM) constitute a type of artificial neural network based on the topological properties of the human brain (Kohonen [7]).

SOM are designed for unsupervised learning and are trained using the Kohonen algorithm. SOM can also be understood as a non-linear mapping of a multidimensional hyperspace onto a two-dimensional lattice (Kohonen [7]). Points which are close to each other in the hyperspace can also be expected to be assigned to the same neuron of in the two-dimensional lattice. SOM have been used for different purposes in air pollution. A thorough revision of SOM applications in air pollution is available in the literature (Hsin-Chung Lu et al. [8]). The purpose in this case was to identify how far points (sensors) were from each other.

2.2 Cluster analysis

Clustering is the assignment of objects into groups. In cluster analysis (CA) cases to be clustered are defined in an *n*-dimensional hyperspace and distances are computed accordingly. Cases that are close to each other are expected to be similar and after a criterion of distance is defined several algorithms can be used to detect groups of cases. In this study, a hierarchical agglomerative algorithm based on single linkage (also known as nearest neighbour) has been used. In single linkage, the distance between two clusters is computed as the distance

between the two closest elements in the two clusters. For this work, the distance chosen has been the Euclidean, with the single linkage criterion.

The sequence of case agglomeration can be described with the linkage distance. The cases (sensors) that are similar tend to agglomerate with each other at small linkage distances, while distant cases in the hyperspace will remain at higher distances.

3 Results

3.1 Self-organizing Maps

Each of the four SO_2 sensors was first represented in a 68-dimension hyperspace as a point and then a SOM was built to project the four sensors (points) onto a two-dimension lattice. The objective was to identify the sensors that in the 68-dimension hyperspace are close to each other. In other words, the aim is to identify which sensors exhibit small spatial variability among them. These sensors can be expected to be assigned to the same neuron in the SOM and as a consequence, the information provided by them could be considered to be redundant.

The 4x1 SOM finally selected, assigned each of the four sensors to different neurons as can be seen in Fig. 2.

This meant that in the 68-dimension hyperspace the four sensors are well far from each other or what is the same, their 68 monthly values are different enough during the analysis period. In case that at least any two sensors had the same overall behaviour, it could have been expected that they were assigned to the same neuron. This was not the case (Fig. 2) so it was concluded that i) there

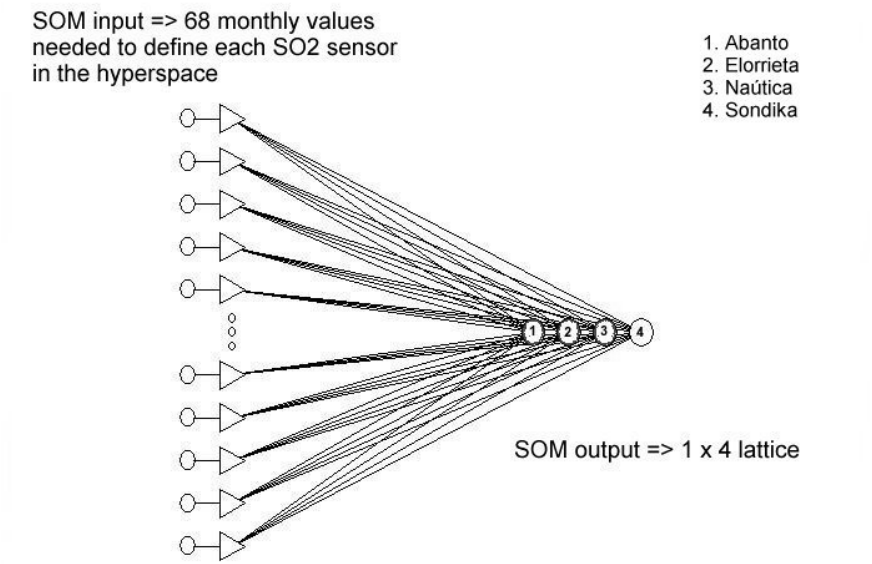

Figure 2: SOM corresponding to the four sensors.

exists an important spatial variability in the SO_2 field of the area, ii) all the information provided by the four sensors was needed to represent this pollutant's field and iii) no significant redundancies were present.

The same as for SOM analysis, each of the four sensors was considered as a point in the 68-dimensions hyperspace. Cluster analysis was carried out using Euclidean distance and the single linkage method applied to the 68 monthly cases of SO_2 in the four sensors. The objective was to identify the sensors that were close to each other.

3.2 Cluster analysis

The results can be seen in Figure 3 (linkage distance during the agglomeration process). Sensors measuring the same and capturing similar SO_2 values and trends during the period analyzed could be expected to link near the bottom.

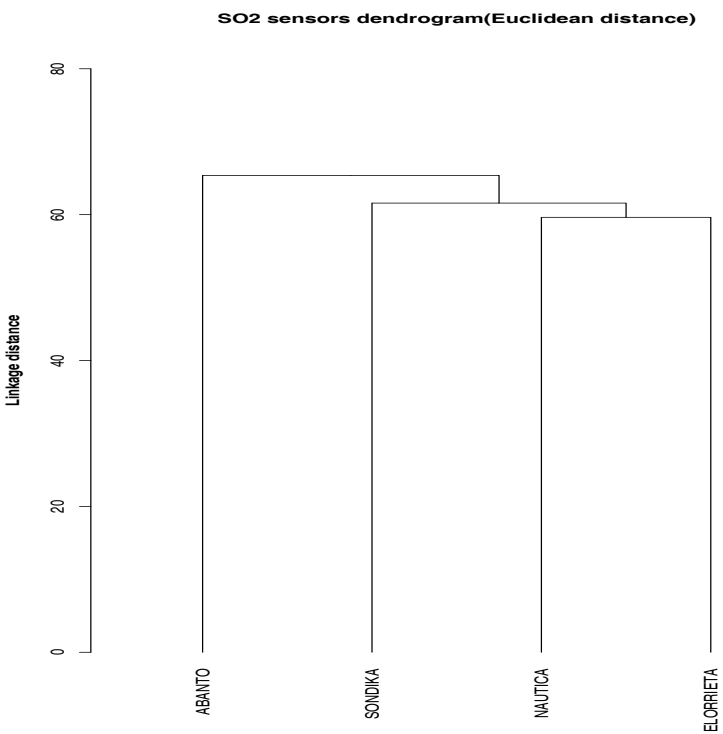

Figure 3: Linkage distance for the four sensors.

This turned out not to be the case and the information that can be obtained from the results of the cluster analysis contained in Figure 3, confirm that the four sensors exhibit notoriously different behaviours, thus confirming the results obtained via SOM.

4 Discussion

The results obtained with SOM and CA agree in that the both techniques lead to the same conclusion that the four sensors do not behave similarly and therefore a significant spatial variability takes place in ZONE B.

The three sensors which do not belong to ZONE A are under the influence of different emission and dispersion patterns which suggests that a box model mechanism is not acting for the whole ZONE B. These results are in agreement with the fact that while in Bilbao city there are a lot of small SO_2 sources, other sensors like Abanto and Sondika are under the influence of nearby SO_2 emission facilities. The orography near these sensors and in the case of Nautica, the nearness to the sea (Fig.1), also originates dispersion mechanisms different from those in ZONE A. This means that emissions from all the sources in the area do not mix perfectly above Great Bilbao and do not get dispersed according to the same mechanisms.

For this reason, it can be concluded that in ZONE B, four different inmission patterns can be distinguished for SO_2.
However, this does not necessarily mean that additional SO_2 patterns not detected yet in the absence of appropriately located sensors, do not exist in ZONE B.

All this indicates a significant spatial variability of the SO_2 field in the area but more sensors might be needed to accurately assess the existing SO_2 field. However, any new configuration will need evaluation and feedback. In this sense, SOM, and CA can be a most valuable tool.

The main advantage of SOM and CA is that their results are readily interpretable and identification of similar/different behaviours in sensors is straightforward. The combination of SOM, and CA yield useful information on redundancies in network configuration and can also be helpful to inspire further changes in the frame of a continuous network optimization process.

5 Conclusions

The methodology presented in this paper has been developed to be applied at a monthly time scale so that it can be used not only with automatic air quality networks but also with networks of passive samplers. A second reason for this has been to eliminate the impact of high frequency cycles which could create a masking effect. SOM, and CA have proved to be independent and complementary tools capable of identifying different behaviours in inmission levels of SO_2 as measured by the sensors. The results show that for Great Bilbao's network the four sensors analyzed are not redundant and all the information provided by them is needed to assess the evolution of the SO_2 field in the area.

Due to the extremely rapid growing rate of many urban areas all over the world (particularly Eastern Asia), air pollution networks also experiment continuous changes in their configuration to cover more and more areas that only a few months earlier were rural. In this context, air pollution networks also need continuous evaluation and feedback, particularly in the aspects of spatial

representation of pollutants. The methodology developed in this paper combines SOM, and CA and though applied to a case study, can be easily applied in those cases where surveillance networks experiment changes in time. If more rapidly reacting pollutants like CO, NOx or ozone were analyzed, the same methodology based upon SOM, and CA could be used to assess spatial variability and evaluate suitability of network configuration regarding spatial representation of these pollutants. However, the much higher reactivity rate of CO, NOx or ozone, involves a study at, not only monthly but also, additional time scales.

This methodology combines cluster analysis and SOM and allows evaluation and feedback of a network's capability in the frame of a continuous network optimization process. Though applied in this work to SO_2 at a monthly time scale, it can be easily applied to any other pollutants (ozone, NOx …), locations and time scales (hourly, daily, weekly, yearly).

In other networks, sensors might not behave so differently as in the present work happens in ZONE B. They might not be so similar to each other as in ZONE A either.

Acknowledgements

This work has been financially supported by the Spanish Ministry of Science and Education (contract CGL2008-03321/CLI-MORECIP National R+D+I Projects) and EKLIMA21 ETORTEK-09 project (Basque Autonomous Government's Meteorological Agency, EUSKALMET). The authors acknowledge the Spanish Ministry of Science and Innovation (CTM2006-02460 financed jointly with FEDER funding) and the University of the Basque Country (Research group GIU08/10UPV) for the financial support for the project. The authors also wish to thank the Basque Government for providing with data for this study.

References

[1] Ibarra-Berastegi, G., Elias, A., Agirre, E., Uria, J., 2001-I. Long-term changes in ozone and traffic in Bilbao. Atmospheric Environment 35, 5581-5592.

[2] Ibarra-Berastegi, G., Elias, A, Madariaga, I., Agirre, E., Uria, J., 2001-II. Short term time forecasting of hourly ozone, NO2 and NO levels by means of multiple linear regression modelling. Environmental Science and Pollution Research 8, 4: 250.

[3] Ibarra-Berastegi, G., Elias, A, Madariaga, I., Agirre, E., Uria, J., 2001-III. Short term time forecasting of hourly ozone, NO2 and NO levels by means of multiple linear regression modelling. Gate to Environmental Health Science 1, 1-7.

[4] Ibarra-Berastegi, G., Elias, A., Agirre, E., Uria, J., 2003. Traffic congestion and ozone precursor emissions in Bilbao (Spain). Environmental Science and Pollution Research 10, 360-360.

[5] Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J., 2008. From diagnosis to prognosis for forecasting air

pollution using neural networks: Air pollution monitoring in Bilbao. Environmental Modelling and Software 23, 622-637.
[6] Zannetti, P., 1990. Air pollution modeling. Theories, Computational Methods and available software. Computational Mechanics Publications, Southampton, UK.
[7] Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59-69.
[8] Hsin-Chung Lu, Chung-Liang Chang, Jen-Chieh Hsieh, 2006. Classification of PM10 distributions in Taiwan. Atmospheric Environment 40, 1452-1463.