Prediction of air pollution levels using neural networks: influence of spatial variability

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

Abstract

This work focuses on the prediction of hourly levels up to 8 hours ahead for five pollutants (SO_2, CO, NO_2, NO and O_3) and six locations in the area of Bilbao (Spain). To that end, 216 models based on neural networks (NN) have been built. Spatial variability for the five pollutants has been assessed using Principal Components Analysis and different behaviour has been detected for the nonreactive pollutant (SO_2) and the rest (CO, NO_2, NO and O_3). This can be explained by the very local effects involved in the photochemical reactions. The inputs used to feed the NN models intended to predict forthcoming levels of these five pollutants, include a baseline based on autocorrelation plus a linear or nonlinear combination of different meteorological and traffic variables. The nature of these combinations is different depending on the sensor thus showing the importance of the spatial variability to build the models. The number of hourly cases, due to gaps in data predictions, can have a possible range from 11% to 38% depending on the sensor. Depending on the pollutant, location and number of hours ahead the prediction is made, different types of models have been selected. The use of these models based on NNs can provide Bilbao's air pollution network originally designed for diagnosis purposes, with short-term, real time forecasting capabilities. The performance of these models at the different sensors in the area range from a maximum value of R^2=0.88 for the prediction of NO_2 1 hour ahead, to a minimum value of R^2=0.15 for the prediction of ozone 8 hours ahead. These boundaries and the limitation in which the number of cases that predictions are possible represent the maximum forecasting capability that Bilbao's network can provide in real-life operating conditions.

Keywords: PCA, neural networks, fluid mechanics, air pollution forecasting, air quality network, traffic network, Bilbao, photochemistry, chemical engineering, applied physics.

WIT Transactions on Ecology and the Environment, Vol 116, © 2008 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR080411

1 Introduction

This work describes the results of a study carried out in the Bilbao area corresponding to years 2000 and 2001, in which data from the three existing networks in this city (air quality, meteorological and traffic) have been analyzed jointly to see if short-term, real time hourly forecasts can be obtained for SO_2, CO, NO_2, NO and O_3.

For the period analyzed in this study (2000–2001) the main sources of SO_2 are small domestic heating systems and to a much lower extent, traffic. CO emissions are mainly due to traffic and, in winter, also domestic heating. As far as NO_2 and NO emissions are concerned, a report to the Basque Government suggests that NOx emissions in the area are mainly due to traffic. All these emissions are scattered throughout the whole area and do not show an important spatial variability.

The underlying assumption for this work has been that if the system formed by the three existing networks can properly describe the joint evolution of air pollution, meteorology and traffic, a thorough analysis of their historical records can detect and recognize patterns and relationships among them and as a result, lead to the prediction of forthcoming air pollution levels. These relationships correspond to well known complicated fluid mechanics and photochemistry mechanisms, which are not always easy to model. However, the links between inputs (current and past values of air pollutants, meteorology and traffic) and outputs (future values of air pollution) can be modelled using statistical techniques. Since the mechanisms involved are known to be highly nonlinear, neural networks have been widely used [1–3]. In this work the effect that spatial variability has on this type of model is analyzed.

2 Database

Data to build and test the models were obtained from the historical records of the three networks existing in the area of Bilbao (Spain) corresponding to years 2000 and 2001. In Fig. 1, the six air pollution sensors are labelled from #1 to #6 and the three meteorological sensors as A, B and C. Hourly values of SO_2, CO, NO_2 and NO are measured at the six sensors in the area while O_3 is only measured at locations #1, #2 and #3.

Temperature and relative humidity are measured at the three meteorological sensors, wind speed only at locations B and C, while atmospheric pressure and radiation only at location A. Since sensor B is nearly at sea level and C is 200 m.a.s.l., the difference of temperature between them can be considered as a descriptive estimator of the true vertical thermal gradient.

Traffic is monitored at 181 locations throughout the central area of Bilbao (black zone in Fig. 1) with sensors located under the streets. At each of them, a variable, which represents the number of vehicles (NV) passing above each sensor every ten minutes, is measured. Hourly averages of NV were calculated and mean hourly values for the whole area were computed using measurements from the 181 sensors.

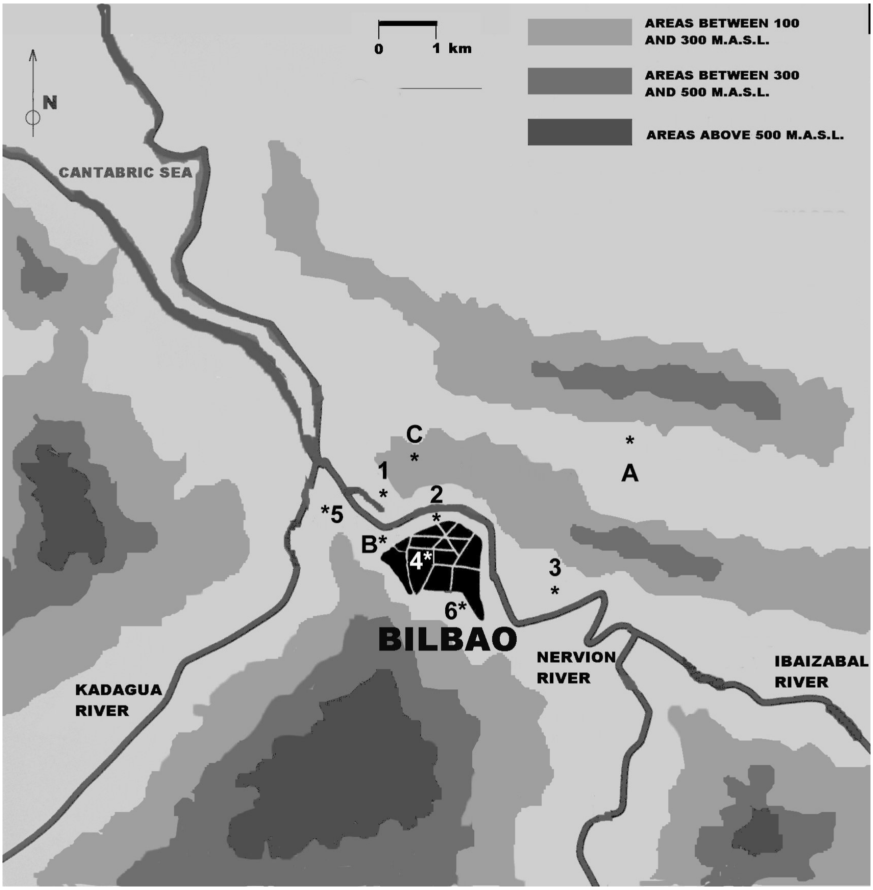


Figure 1: Bilbao and surrounding area.

Finally, all these variables exhibit daily, weekly and yearly cycles, so the sine and cosine functions of these periodicities were calculated and considered as additional candidate inputs. Apart from traffic, additional unknown emission sources were expected to show similar periodicities, so the sine and cosine functions could also be understood as surrogate input variables associated to these unknown pollutants' emissions.

For this study, historical hourly records of the above mentioned variables corresponding to years 2000 and 2001 were available [4–7]. Data of year 2000 were used to build the different groups of candidate prognostic models while data belonging to year 2001 were reserved to test and select the best model. Each case consisted of a complete set of values corresponding to the output and candidate input variables. In principle, for year 2000, 8760 hourly cases with data from the three networks were available. However, due to the gaps and missing data existing in the historical records and also those produced after data pre-processing, the total number of cases available of year 2000 ranged from

1496 in sensor n#2 to 4372 in sensor n#6 and for year 2001, (test set) ranged from 971 in sensor n#2 to 3337 cases in sensor n#5.The goal of this work was to build a group of statistical prognostic models to forecast SO2, CO, NO2, NO and O_3 hourly levels at six locations (Fig. 1) in the city of Bilbao (Spain). The statistical tools employed have been several types of neural networks (NNs). These NNs obtained can be easily incorporated into the air pollution network management activities to obtain hourly forecasts [8,9,10]. The NNs are the core of the BISTAPOF (**BI**lbao **S**hort-**T**erm **A**ir **PO**llution **F**orecast) model. A demo of BISTAPOF is available at no cost from http://www.ehu.es/eolo/software/bistapof_demo/index.html

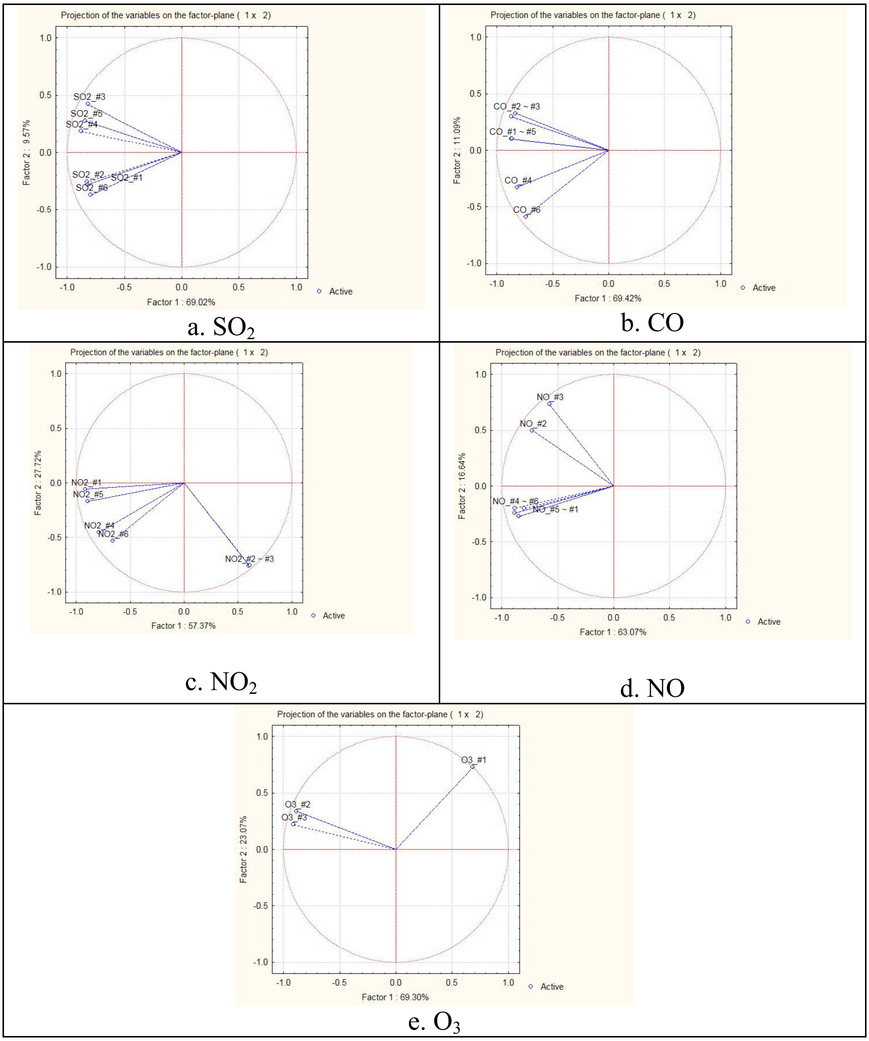
3 Results

For the five pollutants analyzed, predictions are made using all the types of NNs [4–10]. However, it was necessary to analyze the effect that the spatial variability of pollutants' emissions have on the different models used at each location.

To that end, the only pollutant that in the space and time frame of this study could be considered inert -SO_2.-was analyzed separately. The rest of pollutants are involved in complex photochemical reactions that are highly site-dependant like the availability of VOCs or the NOx/VOC ratio and as a consequence, they can be expected to have a higher spatial variability

The tool used to detect spatial variability in the emission fields of SO_2, CO, NO_2, NO and O_3 measured at each of the six sensors during years 2000 and 2001 was Principal Component Analysis (PCA). For the five pollutants, the results of the PCA show that 2 factors are enough to account for fractions of the overall variability ranging form 79% to 92%. For each pollutant, if the measurements in each sensor are represented on the factor plane corresponding to the two main factors, it can be graphically detected clusters of sensors with similar factor loads and subsequently, little or non-significant spatial variability. Inversely, sensors with different factor loads will appear in the graph far from each other, thus suggesting relevant spatial variability (Fig. 2a-2e). In the case of SO_2, the representation of the six sensors on the two-factor plane (Fig. 2a; ~80% of the overall variability) shows similar factor loads, and the sensors tend to cluster close to each other. Therefore, it can be concluded that the six sensors are capable of capturing the same main SO_2 regimes in the area.

If the emissions are scattered throughout the whole area and the SO_2 sensors "see" nearly the same, it can be concluded that the dispersion in the area of Bilbao covered by the six sensors of a non-reactive pollutant like SO_2, can be described following the single box model where the SO_2 is almost perfectly mixed in the boundary layer throughout the whole area. The single box model is based on the mass conservation of pollutants inside a Eulerian box. In the area above the 6 sensors the two mountain ranges (Fig. 1) form a box in which measured SO_2 levels are due to emissions inside the box plus the transportation from or to nearby areas (advection). These SO_2 apportions are associated to the main circulations in the area which take place forth and back along the river axis (SE-NW).



a. SO_2

b. CO

c. NO_2

d. NO

e. O_3

Figure 2: a–e. Factor plane for the five pollutants and six sensors.

In the case of CO, emissions are primarily originated by traffic throughout the whole area and, although involved in the photochemical reactions, CO emission levels are mainly associated to traffic cycles. The traffic flows in the whole area are similar, with low spatial variability, so emissions can also be expected to be similar. The results of the PCA applied to the six CO sensors show quite a similar behaviour (Fig. 2b), which is in agreement with the emission pattern originated by traffic. However, the representation in the factor plane shows that sensors #2 and #3 are particularly near to each other. NO_2 and NO emissions are also mainly due to traffic [4,8,10] and these emissions do not show strong spatial variability. However, both pollutants are involved in photochemical reactions that are highly site-dependant. The representation of NO_2 and NO measured levels on the factor plane (Fig. 2c-2d) shows that again, sensors #2 and #3 tend to behave differently from the rest. Ozone levels are highly site-dependant, and apart from precursor emissions they also depend on the availability of VOCs and the NOx/VOC ratio in the vicinity of each sensor. Sensors #2 and #3 are located near green areas where the local availability of VOCs produced by the vegetation is substantially higher than that in sensor #1. This explains that the ozone production/destruction regimes can be expected to be significantly different in sensors #2 and #3 if compared with those of sensor #1. The representation in the factor plane (Fig. 2e~92% of the overall variability) is in agreement with this and strong spatial variability can be detected between sensor #1 on the one hand, and sensors #2 and #3 on the other hand. Being the physical distance between sensors #2 and #1 smaller than between sensors #2 and #3 (Fig. 1), the question might arise about why sensor #2 shows more similarities with sensor #3 than with sensor #1. This includes higher mean ozone levels in sensor #2 (37.33 $\mu g/m^3$), and #3 (34.65 $\mu g/m^3$) – both above the area's average value of 33.4 $\mu g/m^3$ (Table 1) – than in #1 (28.32 $\mu g/m^3$), below the average The explanation is that the spatial variability mentioned above is not related to the physical distance between sensors, different levels of solar radiation, or different NOx emission patterns near each sensor, but to very local effects like the availability of VOCs and as a consequence, different NOx/VOC ratios leading to different photochemical patterns.

If the two-factors representation for the reactive pollutants (CO and mainly NO_2, NO and O_3) are compared, it can be seen that sensors #2 and #3 tend to behave differently from the rest. This is in agreement with the fact that in the vicinity of sensors #2, and #3 the availability of VOCs can be expected to be higher due to the emissions from green areas nearby. As a result, the photochemical reactions follow different patterns.

Six major types of inputs were identified in the 216 NNs finally selected for the prediction at a given hour H of a certain pollutant's levels H+K=(1,…8) hours ahead:

Type 1: The pollutant's levels measured at current hour H. The information contained in this type of inputs is that of the autocorrelation function at a lag of value K hours.

Type 2: The pollutant's levels measured at hour H-Z, being Z the number of hours before H that the variable has been measured.

Type 3. In the case of pollutants involved in photochemical reactions, the rest of reactive pollutants – measured at hour H or H-Z- usually also appear as inputs.

Type 4: Traffic measured at time H and H-Z.

Type 5. Meteorological variables, mainly radiation and wind speed at time H and H-Z.

Type 6. Sine and cosine functions corresponding to the daily, weekly and yearly cycles.

Very often the typical value of Z =24-K, thus indicating the strength that the 24 h cycle has on all the forecasts. A sensitivity analysis applied to the 216 NNs showed that in all cases, the inputs belonging to type 1 were the most relevant to explain changes in the outputs. This indicates that predictions are made using a baseline, which is the autocorrelation function, that is, current values of the pollutants measured in the network (input type 1). In some cases (31.9%) this is the best option (persistence) while in others, the information corresponding to additional input variables (types 2-3-4-5-6) is also incorporated and combined in the frame of linear (51.4%) or nonlinear (16.7%) models. The rather high amount of models in which persistence is not outperformed indicates that although in some of these cases, linear or MLP models perform equally, there is no need to select a complicated model if a simple one (like in this case, persistence) is enough [10].

In the case of SO_2 (the only non-reactive gas in the frame of this study's time and space scale), persistence is enough in as many as 68.8% of cases. It can be seen that for this pollutant, up to 4h ahead, in most sensors persistence is the best option. From 5h ahead onwards, at least in half the sensors more elaborated models (linear or not), which incorporate the rest of input types, are used. This suggests that being SO_2 a pollutant whose emission levels are closely linked to emissions, four hours can be understood as the average period of time needed to detect changes in emissions and/or transportation from/to nearby areas. During nightly hours, perhaps persistence could be the best option for more than 4 hours ahead. However, in other periods of the day changes in emissions take place more rapidly and the persistence model might not work so well. However, the same model is used for daily and nightly hours so 4 hours can be considered as an average period for the validity of the persistence models in the case of SO_2.

For the CO predictions up to 2 hours ahead, persistence is the best option in four sensors. This gas is reactive but its emission levels are mainly guided by the evolution of traffic. The mean autocorrelation function for traffic up to 2-4 hours lag, shows quite high values which explains why persistence (based exclusively on type 1 inputs), is the best option. Performance of persistence models was aided by the prevalence of moderate CO levels in the area. After 4 hours ahead more elaborated models (linear or not) incorporating additional information corresponding to the rest of input types are needed to predict CO levels.

For the rest of reactive pollutants like NO_2, NO and ozone, autocorrelation (persistence) cannot be used beyond 1h-2h ahead and in some cases, like ozone, not even that. The reason is that concentrations are continuously varying, not only due to changes in the emissions but also owing to their participation in the photochemical processes. Therefore, predictions need to be calculated incorporating (in a linear or nonlinear way) additional information corresponding to the rest of inputs.

The spatial variability described above for reactive pollutants is not reflected in the selected model's architecture but mainly in the relative relevance that the six types of inputs have on the final prediction. A sensitivity analysis shows that, in sensors #2 and #3, the most relevant inputs belong to type 1, followed by type 3 and type 5. In the rest of sensors, after type 1 inputs, it is type 2 and type 4 inputs that contribute most to build the predictions. As said before, sensors #2, and #3 are affected by local emissions of VOCs and the information that needs to be added to the prediction baseline (input type 1) must also include the past behaviour of the rest of reactive pollutants (input type 3) and meteorology (input type 5). In the rest of sensors, there is a major single source of precursors (traffic) and therefore, 24h cycle corresponding to the different pollutants, constitutes the most powerful signal after autocorrelation.

4 Conclusions

Air quality networks are usually designed for diagnosis purposes, being a key feature of a good network that it has enough time and space resolution to follow the evolution of the most important pollutants.

The air pollution network of Bilbao was originally designed as a diagnosis tool to describe in real time the evolution in the area of several pollutants and meteorological variables. The traffic network was also intended to follow the evolution of the traffic flow in the area of Bilbao. Bringing together the information from these networks, statistical models based on NNs to obtain short-term forecasts of air pollution levels can be built. The use of these models can provide the air pollution network with new forecasting capabilities. Very local effects have a great influence in the mechanisms of production for reactive pollutants. This results in an important spatial variability though the NNs can capture this variability and performance is not affected by this fact.

Acknowledgements

This work has been financially supported by the University of the Basque Country UPV-EHU (Spain) (contract #1/UPV 00149.345-E-15398/2003 and research group GIU05/12), the Spanish Ministry of Science and Education (contract CGL2005-06966-C07-05/CLI-INVENTO National R+D+I Project) and project EKLIMA21 ETORTEK (Basque Autonomous Government's meteorological agency, EUSKALMET).

The authors also wish to thank the Basque Government and Bilbao's Municipality for providing data for this study.

References

[1] Gardner, M.W., Dorling, S.R. Artificial Neural Networks (The Multilayer Perceptron). A review of applications if the atmospheric sciences. *Atmospheric Environment,* **32**, pp. 2627–2636, 1998.

[2] Gardner, M.W., Dorling, S.R. Neural network modelling and Prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment,* **33**, pp. 171–176, 1999.

[3] Gardner, M.W., Dorling, S.R. Statistical surface ozone models: an improved methodology to account for nonlinear behaviour. *Atmospheric Environment,* **34**, pp. 21–34, 2000.

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

[5] Ibarra-Berastegi, G., Madariaga, I., Agirre, E., Uria, J. Short-term real time forecasting of hourly ozone, NO_2 and NO levels by means of multiple linear regression modelling. *Environmental Science and Pollution Research,* **8**, pp. 250, 2001

[6] Ibarra-Berastegi G., Elias, A., Agirre, E., Uria, J. Traffic congestion and ozone precursor emissions in Bilbao (Spain). *Environmental Science and Pollution Research,* **10**, pp. 361–367, 2003.

[7] Elías, A., Ibarra-Berastegi. G. Arias, R., Barona, A. Neural networks as a tool for control and management of a biological reactor for treating of hydrogen sulphide. *Bioprocess & Biosystems Engineering*, **29**, pp. 129–136. 2006.

[8] Ibarra-Berastegi, G. Short-term prediction of air pollution levels using neural networks. Air Pollution XIV. pp. 23–32. ISBN 1-84564-165-5. WIT Press. Southampton. UK. 2006.

[9] R. Arias, A. Barona, G. Ibarra-Berastegi, I. Aranguiz and A. Elías. Assessment of metal contamination in dredged sediments using fractionation and Self-Organizing Maps. *Journal of Hazardous Materials* **151**, 78–85. 2008

[10] G Ibarra-Berastegi, Ana Elias, Astrid Barona, Jon Saenz, Agustin Ezcurra, Javier Diaz de Argandoña. From diagnosis to prognosis for forecasting air pollution using neural networks: air pollution monitoring in Bilbao (Spain). *Environmental Modelling & Software*. **23,** pp. 622–637, 2008

[11] A. Ezcurra, J. Sáenz, G. Ibarra-Berastegi, J. Areitio. Rainfall yield characteristics of electrical storm observed in the Spanish Basque country area during the period 1992-1996. *Atmospheric Research. Accepted,* 2008.