

Forecasting high SO₂ concentration levels with fuzzy clustering techniques

J. Ricardo Pérez-Correa[†], M. Victoria Letelier[†], Aldo Cipriano[‡], Héctor Jorquera[†], Oscarina Encalada[†] & Iván Solar[†]

[†]Depto. de Ingeniería Química y Bioprocesos, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile

[‡]Depto. de Ingeniería Eléctrica, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile
Email: perez@ing.puc.cl

Abstract

The city of Copiapó in Chile, close to the Hernán Videla Lira smelting copper plant, often suffers episodes of high SO₂ concentration levels that seriously affect the well-being of its inhabitants. A system forecasting high SO₂ levels was developed in order to alert the people of Copiapó and to give the plant engineers time to alter the operating conditions and thereby prevent the occurrence of these episodes or at least minimise their magnitude. The system uses a fuzzy clustering technique to process on-line data streaming from five monitoring stations every minute. The system is capable of forecasting SO₂ levels precisely two hours in advance during the summer season, when SO₂ concentrations are usually low. However, during winter, when SO₂ levels are high, the system underpredicts maximum values.

1 Introduction

The degree to which public health is impaired by airborne pollutants is highly correlated to the extent in which levels of observed contaminants exceed existing air quality norms. In Chile, USEPA guidelines have been

adopted by the Ministry of Health for the following pollutants¹: total suspended particles, inhalable particulate (PM10), sulphur dioxide (SO₂), carbon monoxide (CO), ozone (O₃) and nitrogen dioxide (NO₂). The Chilean norm for SO₂ is 80 (μg/m³) average annual concentration and 365 (μg/m³) average daily concentration².

The city of Copiapó is affected by high levels of SO₂, caused by the activities of the Hernán Videla Lira/ENAMI copper smelting plant. Local authorities have declared this zone saturated, meaning that no additional sources of SO₂ emissions are allowed. An advance warning system is desirable in order to minimise the effects of these pollutants on the population. One that would forecast peak values enabling source emissions to be reduced before severe pollution affects public health.

The use of advanced identification techniques is necessary in the development of reliable forecasting systems. Several studies evaluating such identification techniques in pollution control problems have been published. Ziomas³ presents various analytic models for predicting maximum NO_x levels using meteorological data. Yi and Prybutok⁴, have used neural nets in order to predict daily maximum ozone concentrations. Boznar et al.⁵ apply neural nets to develop an environmental forecasting system in Slovenia. Jorquera et al.⁶ evaluate three identification techniques for forecasting daily maximum ozone concentration a day in advance, using only temperature and the level of ozone as input variables. A fuzzy clustering identification algorithm proved the most promising technique.

The objective of this work is the assessment of fuzzy clustering techniques in the identification of dynamic models for environmental problems that include many input variables. Specifically, this technique will be evaluated in the design of a system for forecasting SO₂ levels in the city of Copiapó, two hours in advance.

2 Fuzzy clustering

Cluster analysis is an established statistical technique, facilitating the classification of a group of elements in terms of their differences or similarities. In order to generate a simple structure starting from a complex group of data, it is necessary to define a measurement base of similarity, which is usually associated with a measure of distance, with the Euclidean distance the most used. Given a measure of similarity, classical

clusters partition the input space assigning each element to a unique cluster.

Fuzzy clustering is used to represent input/output relationships where input variables (fuzzy variables) may belong to several sets (clusters) simultaneously, and where a different input/output model is associated with each cluster. Fuzzy cluster models are expressed in terms of conditional sentences (rules) of the type "IF THEN", containing premises and consequences:

IF (list of premises) THEN (list of consequences) (1)

Here, each element is assigned a degree of membership to each cluster, resulting in a fuzzy partition of the space. This enables fuzzy clustering to attain a better representation of the structure of the input space. An identification algorithm for fuzzy cluster models contemplates two stages. Initially, the degrees of membership and the number of clusters is optimised, such that they satisfy given grouping criteria. This generally produces a non linear programming problem. In the second stage, the parameters that define the models of the consequences, should be adjusted. One input/output model is required for each cluster, and in most of the cases these are linear multivariable models. If this is the case, the parameters can be adjusted by least square algorithms.

3 Identification algorithm

3.1.- Degrees of membership and optimal number of clusters

The standard FCM⁷ (Fuzzy Clusters-Means) method is used here. First, the minimisation of the sum of the fuzzy distances of all the samples, is required, for an arbitrarily given number of clusters. This is the distance from a sample to a given cluster, weighted by the degree of membership of the sample to that cluster. Mathematically, this is represented by:

$$\text{Min } z_{\alpha}(U, C) = \sum_{i=1}^{N_c} \sum_{k=1}^m (\mu_{ik})^{\alpha} d_{ik}^2 \quad (2)$$

subject to:

$$\sum_{i=1}^{N_c} \mu_{ik} = 1; \quad \forall k = 1, \dots, m \quad (3)$$

where,

μ_{ik} : degree of membership of the k sample to the cluster i .

d_{ik} : distance from the k sample to the centre of cluster i .

α : fuzzyfication parameter (heuristic).

N_c : number of clusters.

m : number of samples.

The fuzzyfication parameter, α , defines how fuzzy the clusters are, i.e., if zero, the clusters are conventional. However, the larger parameter value the fuzzier the clusters will be. The recommended value for α is 2, which also assures convergence of the algorithm. From the above minimisation, the centres of the clusters and the degrees of membership of the samples to the respective clusters, are obtained.

The following approach is used for optimising the number of clusters in the input space: "there should be as great a difference as possible between the clusters, but as little difference as possible within them". Mathematically, this means that variance between clusters should be maximum, and the variance within clusters minimum, which is given by the minimisation of the following objective function:

$$S(N_c) = \sum_{i=1}^{N_c} \sum_{k=1}^m (\mu_{ik})^\alpha \left(\left\| x_k - c_i \right\|^2 - \left\| \bar{x} - c_i \right\|^2 \right) \quad (4)$$

where,

x_k = a given sample k .

—

$\bar{x}$ = average vector of all the samples.

—

c_i = centre of cluster i .

—

As the $S(N_c)$ function has been built in such a way so as to eventually reach a minimum as N_c grows, the process produces the optimum number of clusters. This is because by maximising the external variance, the number of elements associated with each cluster diminishes, thereby reducing the internal variance.

3.2 Parameters of the consequences

These are identified using the Takagi-Sugeno⁸ method, where the input/output relationship of each cluster could be represented by a linear combination of the inputs plus a constant. In a model with several inputs and one output, which is the one of interest here, the rules are of the form:

$$R^i: \text{if } \underline{x} \text{ is } A^i \text{ then } y^i = \beta_o^i + \sum_{j=1}^p \beta_j^i x_j \quad (5)$$

where,

$\underline{x} = \{x_1, x_2, \dots, x_p\}$: vector of input variables, of dimension p.

$\underline{A}^i = \{A^i_1, A^i_2, \dots, A^i_p\}$: vector, of p dimension, of the membership functions of the samples to the clusters.

$\underline{\beta}^i = \{\beta_o^i, \beta_1^i, \dots, \beta_p^i\}$: vector of dimension p+1 that corresponds to the parameters to be identified for the ith rule.

y^i : output according to cluster i.

The final estimated output value, is computed as a weighted average of N_c fuzzy outputs, this is:

$$y = \frac{\sum_{i=1}^{N_c} [w^i(\underline{x}) \cdot y^i]}{\sum_{i=1}^{N_c} w^i(\underline{x})} \quad (6)$$

where $w^i(\underline{x})$ corresponds to the activation degree of the ith rule for $\underline{x}$, computed according to:

$$w_0^i = \text{Min}\{\mu_{iA_1}(x_{10}), \mu_{iA_2}(x_{20}), \dots, \mu_{iA_p}(x_{p0})\} \quad (7)$$

A least squares algorithm minimising the the average quadratic error, e^2 , adjusts the parameters of the consequences, with:

$$e^2 = \frac{\sum_{k=1}^m (y_k - \hat{y}_k)^2}{m} \quad (8)$$

where m represents the number of samples used in carrying out the identification, and y_k and $\hat{y}_k$ the measured and inferred output for sample k , respectively.

4 Results and Discussion

4.1 Performance indices

The quality of the model is evaluated through comparison between the observed data, given by the Copiapó monitoring station, and the predicted values. Several indices may be used in assessing the quality of a model. In this work, the following were used:

4.1.1 Square Root of the Mean Square Error (I1)

$$ERMS = \left[\frac{1}{N} \sum_{i=1}^N (o_i - p_i)^2 \right]^{1/2} \quad (9)$$

where,

o_i : observed value on time i .

p_i : predicted value on time i .

N : total number of data.

A small value of this index indicates a good quality model.

4.1.2 Index of Agreement (I2)

$$IA = 1 - \frac{\sum_{i=1}^N (o_i - p_i)^2}{\sum_{i=1}^N (|p_i'| - |o_i'|)^2} \quad (10)$$

where,

$$p_i' = p_i - o_{avg}$$

$$o_i' = o_i - o_{avg}$$

$$o_{avg} = \frac{1}{N} \sum_{i=1}^N o_i, \text{ observed average value.}$$

The index of agreement is an adimensional ranging from 0 to 1, and is a normalised measure of the mean square error. A value of I2 near 1 indicates close agreement between the observed and predicted values. Contrarily, a value of I2 near zero is indicative of a poor model.

4.2 Calibration and validation

There are five monitoring stations around the city of Copiapó. Four of these stations (Paipote, Tierra Amarilla, San Fernando and Copiapó), provide wind speed and direction, and SO₂ concentration data sampled every 60 seconds. The fifth station, which is meteorological only, delivers pressure, temperature, relative humidity, solar radiation, and wind speed and direction values, also measured each minute. The model was calibrated using 55 days of data, from December 1993 to January 1994, when variables were available hourly. FCM was used to optimise the number of rules (clusters), and Least Squares to adjust model parameters. MATLAB was employed in both cases. Five rules proved optimal for this model, producing an average error (I1) of 2.31% between the predicted and observed values.

Validation of the model was carried out using data over the following 32 hours, forecasting SO₂ levels in Copiapó two hours in advance. The result can be seen in the following figure:

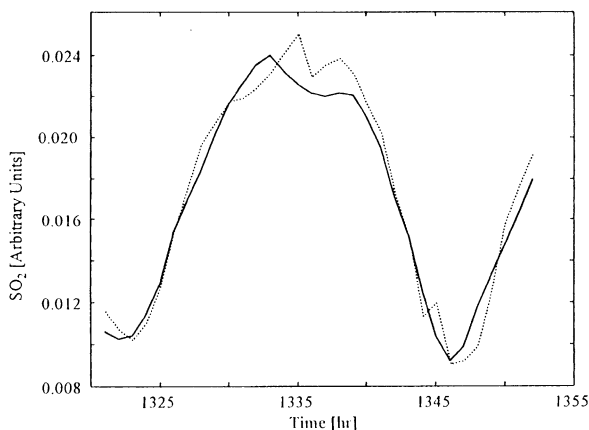


Figure 1: SO₂ forecast in Copiapó during the summer of 1994/1995.
(—) observed value, (.....) predicted value.

The error index I1 in this case was $9.6 \cdot 10^{-4}$ %, and the index of agreement (I2) was 99%, indicating a very good prediction, as seen in the

figure. At first glance, given the values of these performance indices, it could be argued that the model is able to forecast well. However, due to low values measured during this period, nothing can be said *a priori* about the ability of the model to forecast severe pollution episodes, characterised by rapid increases in levels of SO_2 .

In order to test the model under more demanding conditions, July data (winter) 1995 were used, in which significantly higher SO_2 levels were registered than during the summer of 1994/1995. Variables were available every minute in this case. Since the measured variables were contaminated with noise, they were filtered by simply time averaging them. Averages were taken each 5, 10, 15 and 60 minutes, evaluating the predictive capacity of the model in each case; Table 1 shows the results obtained. The calibration set includes data over 9 consecutive days and validation was carried out with the data of the following day. Again, as with summer data, the number of optimum clusters was five.

From this table it can be observed that the model forecasts better for longer filtering periods. Figure 2 shows the results upon using samples with data filtered every 60 minutes.

Table 1: Forecasting results for different filtering periods.

Filtering period [min]	I1($\mu\text{g}/\text{m}^3$)	I2(%)
1	202	23
5	152	30
10	149	31
15	93	71
60	59	88

The figure shows that the model is able to warn the episodes, but it cannot predict their magnitude accurately. This aspect could be improved using a larger set of calibration data incorporating several pollution episodes, or by using a recursive least square algorithm to estimate the parameters of the consequences on-line, enabling the model to be permanently updated.

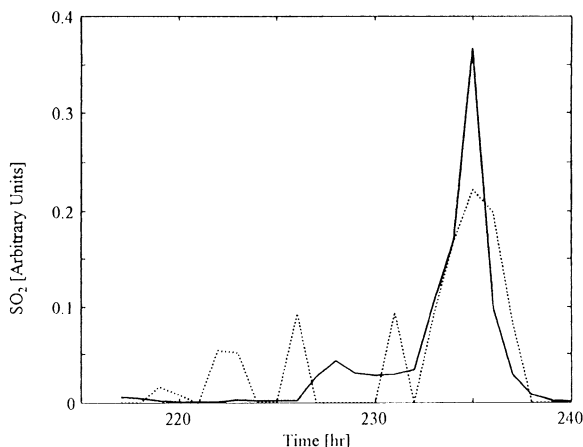


Figure 2: Forecasting SO₂ in Copiapó during the winter of 1995.
(—) observed value, (.....) predicted value.

6 Conclusions

Using the procedure described, it is possible to forecast SO₂ levels in the city of Copiapó well during the summer, two hours in advance, when there are low levels of contamination. In this test, an Index of Agreement of 99% was obtained with the validation set.

Forecasting SO₂ concentration during winter, characterised by frequent episodes of pollution, is more difficult with the model not performing as well. The model is however, able to forewarn of an episode, but can not predict its magnitude accurately. Even so, this should still prove sufficient in order to design a prevention strategy involving modifying plant operating conditions that would attenuate the magnitude of the episode. Moreover, under these conditions the model's performance could be improved by incorporating a more extensive data base or by using a recursive least squares algorithm to adjust model parameters on-line.

References

- [1] Escudero, J. Situación global calidad de aire en la región metropolitana, *Apuntes de Ingeniería*, **19**, N° 3, pp. 7-19, 1996. (in spanish)



- [2] CONAMA-Metropolitana, *Inventario de Emisiones Atmosféricas*, Comisión Nacional del Medio Ambiente-Región Metropolitana, 1993.
- [3] Ziomas, I. Forecasting peak pollutant levels from meteorological variables, *Atmos. Environ.*, **29**, N° 24, pp. 3703-3711, 1995.
- [4] Yi, J. & Prybutok, R. A Neural Network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, *Environ. Pollution*, **92**, N°3, pp. 349-357, 1996.
- [5] Boznar, M., Lesjak, M. & Mlakar, P. A Neural network-based methods for short-term predictions of ambient SO₂ concentration in highly polluted industrial areas of complex terrain, *Atmos. Environ.*, **27B**, N°2, pp. 221-230, 1993.
- [6] Jorquera, H., Pérez-Correa, R., Cipriano, A., Espejo, A., Letelier M.V. & Acuña, G. Forecasting ozone daily maximum levels at Santiago, Chile, *Atmos. Environ.*, to be published, 1998.
- [7] Hartigan, J., *Clustering Algorithms*, John Wiley & Sons, New York, 1975.
- [8] Takagi, T. & Sugeno, M. Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. Sys., Man and Cyb.*, **SMC-15**, N°1, pp. 116-132, 1985.

Acknowledgements: This work was supported by projects FONDEF 2-48, FONDEF PTT-009/96, FONDECYT 198-0972 and FONDECYT 196-0394. A criticism by Alex Crawford is appreciated.