PREDICTION OF CARBON MONOXIDE (CO) ATMOSPHERIC POLLUTION CONCENTRATIONS USING METEROLOGICAL VARIABLES

IGNACIO J. TURIAS[1]*, JOSÉ M. JEREZ[2], LEONARDO FRANCO[2], HÉCTOR MESA[2],
JUAN J. RUIZ-AGUILAR[3], JOSÉ A. MOSCOSO[3] & MARÍA J. JIMÉNEZ-COME[3]
[1]Department of Computer Science Engineering, University of Cádiz, Spain
[2]Department of Computer Science, University of Málaga, Spain
[3]Department of Civil and Industrial Engineering, University of Cádiz, Spain

ABSTRACT

This study proposes a two-stage procedure to better predict carbon monoxide (CO) concentrations in the Bay of Algeciras (Spain). In the first stage, a multiple regression model was employed to predict CO concentrations in different monitoring stations using historical data. In the second stage, a new regression scheme was used to forecast the CO concentrations using historical data together with weather forecasts. The experiment shows that two-stage models outperform the single models in the CO concentrations forecasting. The application of the proposed technique may become a supporting tool for the prediction of the values of CO concentrations in complex scenarios such as Bay of Algeciras (Spain).

Keywords: forecasting, air pollution, regression models, resampling procedure.

1 INTRODUCTION

The alarmingly increasing pollution levels in air has been attracting public attention, since it can cause serious health problems. The United States Environmental Protection Agency (EPA) sets carbon monoxide (CO) lead as one of a set of six 'critical pollutants' [1]. Also, EU environment law covers aspects in air and water quality, greenhouse gases and toxic chemicals. The EU promotes the environmental concern and is a major global force in pushing for tighter environmental standards [2].

Urban air pollution has been considered as a local problem mainly associated with urban conditions and industrial emissions. Nowadays, urban environments are, in general, dominated by traffic emissions [3]. The main traffic-related pollutants are CO, NO*x*, hydrocarbons, and particles. CO comes from the imperfect fuel combustion products such as gas, coal or wood. Atmospheric pollutants are responsible for different effects on human health (chronic and acutes). CO is a colourless, odourless, non-irritating but very poisonous gas. CO enters the bloodstream through the lungs and reduces the ability of blood to carry oxygen to the different organs and tissues [4]. Therefore, short-term exposure to high CO concentrations might cause an acute health impact (in small amounts, it causes drowsiness or slow reflexes). If concentrations are sufficiently high, CO may cause death [5].

It is known that air pollution has direct effects on human health through exposure to high concentration of ambient pollutants. Then, air pollution control and the associated prediction of pollutant levels are needed to take preventive and evasive actions. Usually, meteorological variables can be predicted by using a traditional autoregression modelling or by using a multiple regression model. Here, one of the aims is to check if the available meteorological variables help to the regression models to better predict of CO concentrations.

ORCID: http://orcid.org/0000-0003-4627-0252

WIT Transactions on Ecology and The Environment, Vol 211, © 2017 WIT Press
www.witpress.com, ISSN 1746-448X (on-line)
doi:10.2495/AIR170141

To assist the control of CO pollution, the EU directives established a maximum CO threshold of 10 mg/m^3.

In this paper, the authors use different regression approaches in order to provide 1-h advance forecasts of the values of CO concentrations in the Bay of Algeciras. Authors have previously developed different studies to predict atmospherical pollutant concentrations [6]–[9]. The objective of this research is to obtain a suitable prediction model that would enable us to predict the values of pollutant concentrations using exogenous information (in particular, weather forecasting).

Different methods have been used in the prediction of atmospheric pollution: persistence models [10] or regression models [11]–[13]. In this work, multiple linear regression (MLR) models, with pollutant concentrations, speed and wind direction, temperature and time interval as inputs in an autoregressive arrangement, were used. Furthermore, different models were built, some of them using only the pollutant information and the rest considering exogenous variables (with/without weather forecasting).

A procedure of resampling simulation was designed to avoid variation coming from different sources, thus guaranteeing independence and randomness [14]. Authors have applied this procedure successfully in previous works [7]–[9]. The results obtained from the different MLR models with autoregressive inputs were statistically analysed and compared. The predictions are significantly promising.

2 DATA AND AREA DESCRIPTION

The available data covers a period of six years between 2010 and 2015. CO concentration data were collected in five monitoring stations (GUADARRANQUE, E. DE HOSTELERIA, CORTILLIJOS, CAMPAMENTO, ALGECIRAS EPS) and meteorological data extracted from a meteorological tower (T.M. CEPSA) located in the Refinery Plant "Gibraltar-San Roque' (see Fig. 1). Meteorological variables were measured at 60 m in height. Descriptive statistical can be found in Table 1. The monitoring stations where CO hourly measures were obtained are controlled by the Environmental Agency of the Andalusian Government. Gaseous pollutants are monitored by chemical analyzers. The calibration process of all the sampling monitors is supervised by the Environmental Agency of the Andalusian Government. This work is part of the coordinated research projects TIN2014-58516-C2-1-R and TIN2014-58516-C2-2-R supported by MICIN (Ministerio de Economía y Competitividad-Spain).

The 'Bay of Algeciras' region is a very industrialized area where very few air pollution studies have been carried out. Up to date, no model has been developed in order to predict air pollutant levels in the monitoring stations spread in the region as a function of the values of the other rest of stations. About 300,000 inhabitants live in the different towns spread in the 'Campo de Gibraltar'. It is a complex industrial scenario, where many stationary sources are present (Fig. 1): an oil-refinery and some petrochemical factories close to it, a coal-fired power plant, a fuel-oil power plant, a large steel factory and also the Gibraltar airport. Traffic is especially concentrated in the urban areas and the main road of the region (N-340) that surrounds the Bay of Algeciras. The port of Algeciras, one of the most important ship-trading ports in Europe, is another possible source of particulate and gaseous air pollution in the area.

The available data was divided into three parts: training data that was used to build up the models, validation data that was used to select the parameters of the models that best perform on these data, and the testing data that was neither used in building the models nor on selecting the models parameters. The results are provided only using test data in order to compare real world performance of the models.

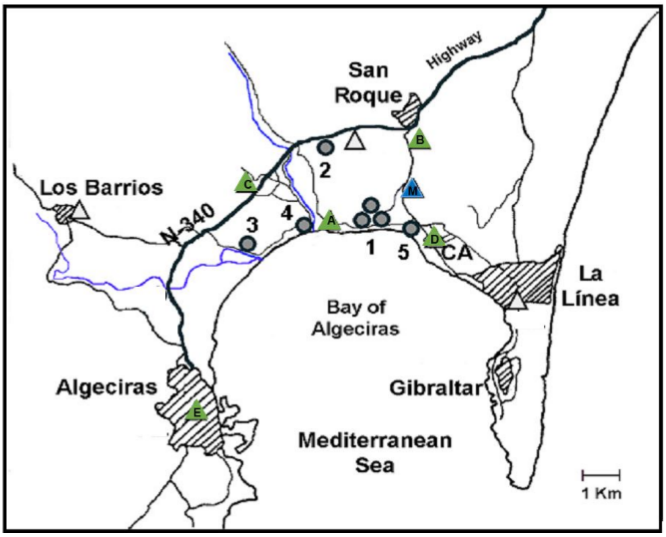


Figure 1: Location of large factories and the monitoring stations (green and blue triangles) in a Bay of Algeciras schematic representation. Stationary sources of pollution are marked with circles: 1) Refinery; 2) Petrochemical factory; 3) Steel factory; 4) Carbon power plant; 5) Fuel oil power plant.

Table 1: Monitoring stations location.

Station	*Abbrev.*	*Lat.*	*Long.*
T.M. CEPSA (M)	TM	36°11’37.66”N	5°24’1.24”O
GUADARRANQUE (A)	GUAD	36°10’55.55”N	5°24’41.06”O
E. HOSTELERIA (B)	EHOS	36°12’13.97”N	5°23’1.33”O
CORTIJILLOS (C)	CORT	36°11’25.74”N	5°26’8.73”O
CAMPAMENTO (D)	CAMP	36°10’45.96”N	5°22’37.09”O
ALGECIRAS_EPS (E)	EPSA	36°8’10.44”N	5°27’12.32”O

Table 2: Descriptive statistics in the monitoring stations (data expressed in μg/m^3).

	EPSA	*CAMP*	*CORTI*	*EHOS*	*GUAD*
Mean	413.884	638.224	555.218	451.386	478.119
Median	391.028	638.264	510.847	433.000	419.278
Mode	313.222	610.139	436.083	412.278	242.139
Deviation	170.105	110.071	206.750	169.423	300.495
Kurtosis	4.089	5.684	6.427	7.899	15.124
Skewness	0.812	0.643	0.773	1.034	2.128
% Gaps	5.830	4.490	6.326	7.558	9.039

3 METHODS

A multiple regression approach has been used in order to predict future CO concentration values in different monitoring stations in the “Bay of Algeciras” region in Spain. The objective is the prediction of 1-h ahead CO values. Three scenarios has been tested to compare the effect of including meteorological variables as inputs. At the first scenario, the CO forecasting at time t+1 is produced using only the past values of CO concentrations values, as eqn (1) shows:

$$\widehat{CO}(t+1) = f\big(CO(t), CO(t-1), \ldots CO(t-n)\big). \tag{1}$$

The second scenario also uses the meteorological variables measured at the monitoring station T.M. CEPSA (see Table 1). The forecasting can be written as eqn (2):

$$\widehat{CO}(t+1) = f\big(CO(t), CO(t-1), \ldots CO(t-n), w_1(t), w_2(t), \ldots w_m(t)\big). \tag{2}$$

where w_i are the available weather meteorological variables (wind speed, wind direction, temperature, …). The third scenario uses the prediction of meteorological variables. Therefore, the model can be formulated as eqn (3):

$$\widehat{CO}(t+1) = f\big(CO(t), \ldots CO(t-n), \widehat{w_1}(t+1), \widehat{w_2}(t+1), \ldots \widehat{w_m}(t+1)\big). \tag{3}$$

For the best results, the data was normalized between 0 and 1 by dividing all the data by the maximum. The training data is used to design the model, while the cross-validation data is used for model selection. The testing data has never been used on building the model and is used to estimate the performance of the network on future unseen data. The designed procedure of random resampling permits an adequate and robust comparison of the tested models comparing the results obtained with test data.

The performance measurements adopted throughout this paper are the root mean square error (RMSE) and the Pearson correlation coefficient (R).

3.1 Multiple regression models

Multiple linear regression (MLR) analysis is a statistical method that allows us to examine how multiple independent variables (x) are related to a dependent (y) variable. Once you have identified how these multiple variables relate to your dependent variable, you can take information about all of the independent variables and use it to make much more powerful and accurate predictions.

Multiple linear regression fits a line (plane or hyperplane) through a multi-dimensional cloud of data points. The general form of the multiple linear regression is defined as eqn (4):

$$y = \sum_{i=1}^{n} \beta_i x_i + \varepsilon. \tag{4}$$

Several requirements such as multicollinearity and homoscedasticity have been checked. There is a number of methods to solve Multiple Linear Regression (MLR) problems [15]: Gauss-Jordan, LU, or QR decomposition because the calculation of the regression parameters by directly inverting matrix X of independent variables could be really dangerous. Nevertheless, the most effective method is Singular Value Decomposition which handles all problems that may arise, such as singularities or ill-conditioning.

4 RESULTS AND DISCUSSION

The experimental procedure was developed using R-software, the popular suite of machine learning software. As we mentioned above, three scenarios have been defined in order to compare if the use of exogenous variables (scenario 2: CO+meteo) outperforms the results of scenario 1 and, also, if the use of predictions of the exogenous variables (scenario 3: CO+meteo2) produces better results. Scenario 1 consists in using only the historical CO data in each monitoring station. Scenario 3 can be considered as a two-stage prediction approach since the model requires as inputs the prediction (done at stage 1) of meteorological variables. In this paper, we have used directly the real values of the future values of the measurements in order to compare purely the effect of introducing these values as inputs.

A resampling strategy with 10-fold cross-validation has been applied, using different quality indexes to evaluate the performance of the prediction models. Two-stage models achieved (in general) better root mean square errors (RMSE) and correlation coefficient (R) quality indexes.

Figs 2–6 indicate that the two-stage model outperforms single model on all orders (different sizes of the past information window). These figures also indicate that, as expected, the more the system order increases, the more the performance of the system improves until a point in which the trend varies and a greater window (using more information) does not improve the results.

Fig. 2 shows how the use of weather forecasting values as inputs improves all the models tested in the monitoring station of GUADARRANQUE. The best values were obtained with a lag (size of lagged values) equal to 3. Using more information does not improve the results and a decrease of the values of the quality indexes can be observed. ALGECIRAS_EPS concentration values presents a very similar behaviour to GUADARRANQUE (Fig. 6). The optimum value of lagged window is obtained at 4. However, the values of R2 coefficient are sensibly lower (R2 = 0.690) than in the case of ALGECIRAS_EPS (R2 = 0.833). Probably, these results can be explained because GUADARRANQUE monitoring station is located very close to a river and a few meters to the shore. Therefore, the local meteorological variables may differ significantly from those measured at TM_CEPSA.

Scenario 3 also presents better results in the case of E. HOSTELERIA monitoring station (Fig. 3). The lag value equal to 9 presents the best values of RMSE and R2 indexes. At the monitoring station of CAMPAMENTO, Fig. 5 shows again how the scenario 3 improves the results. In this case, the value of the lagged values that optimizes the quality indexes is obtained at 5 hours.

Fig. 4. shows that the results are very similar between the three scenarios in the monitoring station of CORTILLIJOS. Probably, these results can be explained due to the fact that this measurement station is located very close to the main road (N-340) of the region.

Additionally, a dependency analysis has been developed. Table 3 shows the dependencies found between meteorological variables and the monitoring stations (measured at the same time *t*). Cross correlation has been calculated in order to find out what are the most relevant variables for explaining each time series of concentration values in each monitoring station. Fig. 7a shows the dependencies in a graph. Dependencies between GUADARRANQUE, CAMPAMENTO and E. HOSTELERIA stations were found. Also, CORTILLIJOS station was linked to GUADARRANQUE and E. HOSTELERIA, and ALGECIRAS_EPS to GUADACORTE. The monitoring station of ALGECIRAS_EPS seems to be the most linked to meteorological variables, although several dependencies between other monitoring stations can be observed.

Additionally, Fig. 7(b) shows the dependencies between monitoring stations using lagged historical data at different times (*t-1*, *t-2*, …). It is worth mentioning that the most relevant lagged dependencies are found using the own information of each station and only weak dependencies can be observed between GUADARRANQUE, CAMPAMENTO and E. HOSTELERIA monitoring stations. These three stations are located closely.

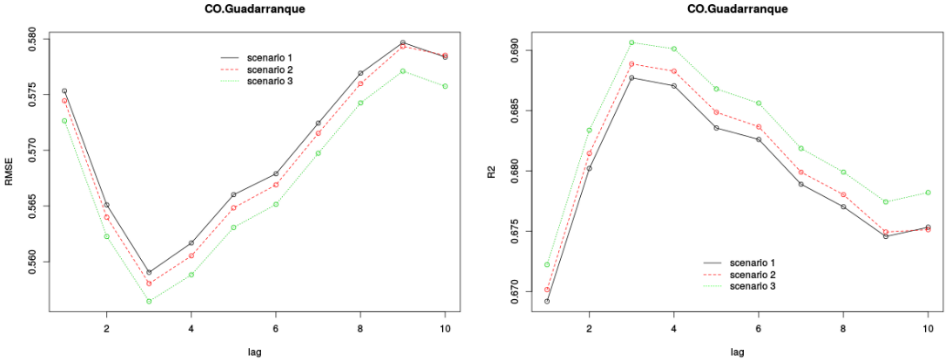


Figure 2: RMSE error and R2 index. Monitoring Station GUADARRANQUE.

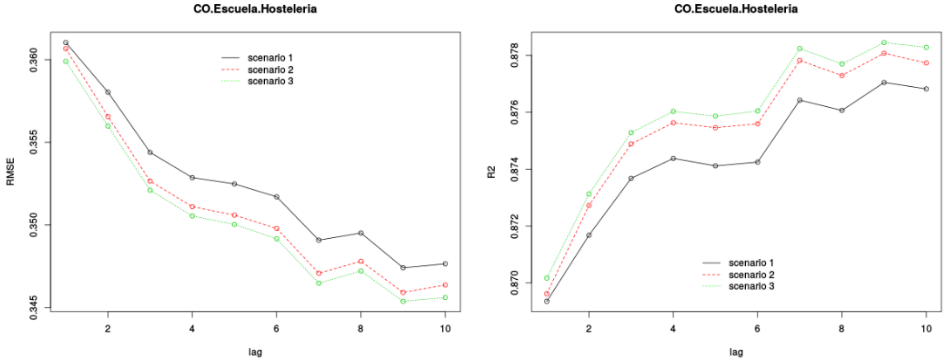


Figure 3: RMSE error and R2 index. Monitoring Station E. DE HOSTELERIA.

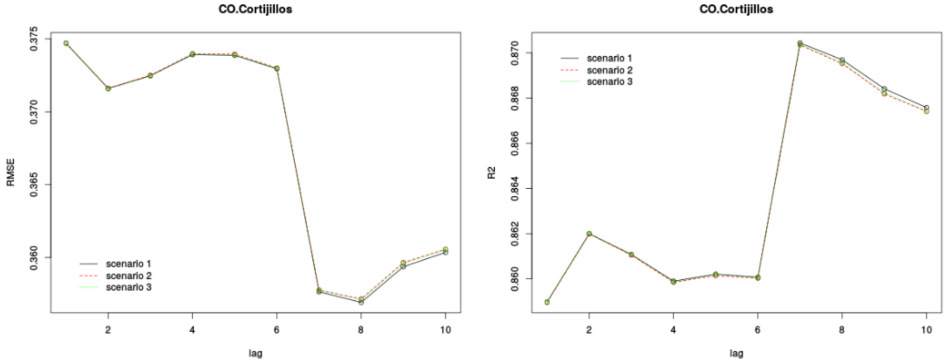


Figure 4: RMSE error and R2 index. Monitoring Station CORTILLIJOS.

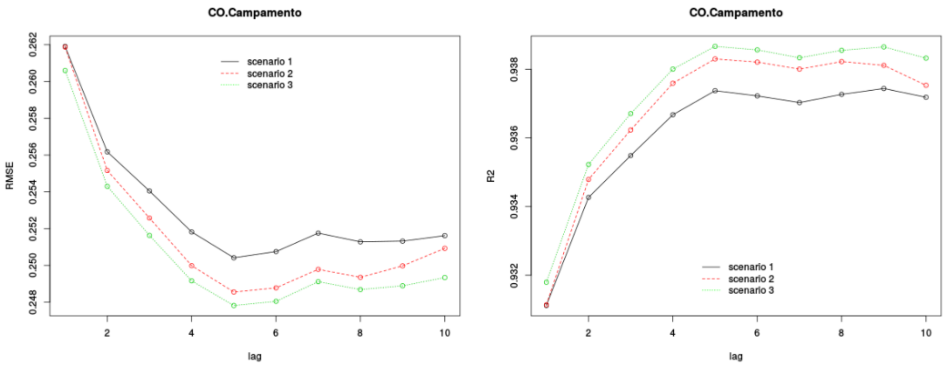


Figure 5: RMSE error and R2 index. Monitoring Station CAMPAMENTO.

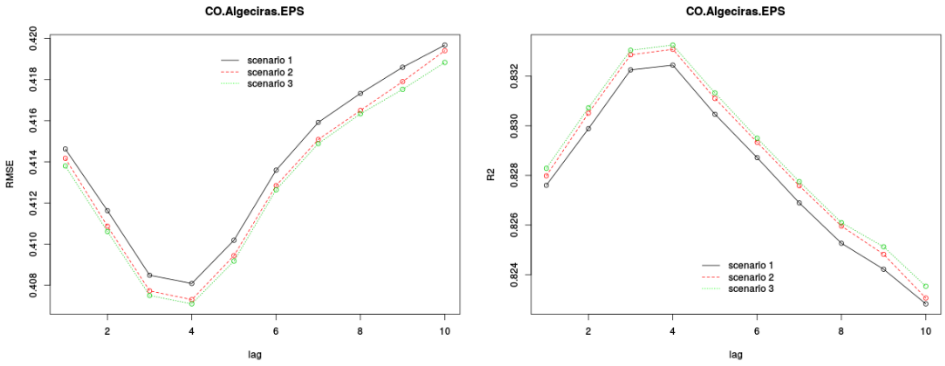
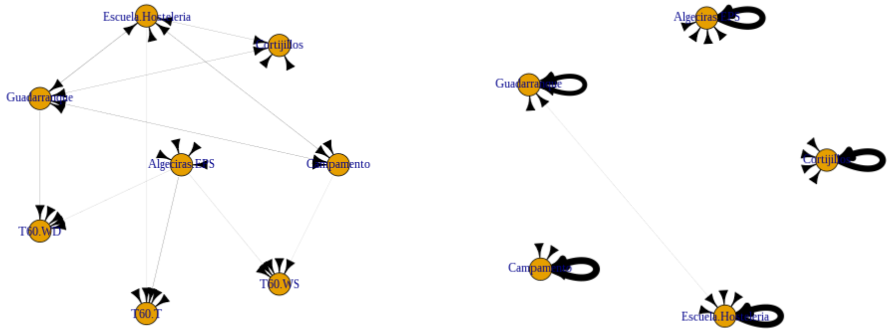


Figure 6: RMSE error and R2 index. Monitoring Station ALGECIRAS EPS.



(a) (b)

Figure 7: (a) Dependency graph between variables and stations (without lagged information); (b) Dependency graph between stations (using lagged information).

Table 3: Dependency values between variables (measured at the same t). T60.WD, T60.WS and T60.T, are respectively, wind direction, wind speed and temperature measured at TM_CEPSA.

	EPSA	*CAMP*	*CORTI*	*EHOS*	*GUAD*	*T60.WD*	*T60.T*	*T60.WS*
EPSA	0.00000	0.0986	0.00846	0.07505	0.64432	0.94139	2.44988	1.00461
CAMP	0.09949	0.0000	0.03234	1.21989	1.07861	0.47703	0.00858	0.71190
CORTI	0.00866	0.0326	0.00000	0.80248	0.86602	0.14891	0.02294	0.06519
EHOS	0.07579	1.2151	0.79021	0.00000	1.45483	0.27481	0.79176	0.00992
GUAD	0.64134	1.0630	0.84309	1.43863	0.00000	1.89266	0.23100	0.41241

The authors evaluated the dependencies between the explanatory variables and the measurements. They improved their model (decreased root-mean-square error and increased R2 index) by incorporating predictions of meteorological data. Results of R2 index values of 0.690, 0.878, 0.870, 0.938, 0.833, respectively for GUADACORTE, E. HOSTELERIA, CORTILLIJOS, CAMPAMENTO and ALGECIRAS_EPS stations for CO(t+1) values showed the ability of the scenario 3 to predict concentration values with better accuracy.

5 CONCLUSIONS

In this paper, modelling of carbon monoxide (CO) concentration using a regression technique is presented. A resampling procedure based on cross-validation is used for parameter identification of the regression models. The approach consists of three scenarios which effectively realize structure identification (inputs) and parameter identification (the lags in the past). The procedure is concretely demonstrated by a simple approach that has been used with the data collected in different monitoring stations in the Bay of Algeciras (Spain). The results show effectiveness of the improved two-stage approach (using weather forecasting). Important technical considerations for selecting the best model include data pre-processing, neighbour relevance and the size of lagged information have been extracted. Results have been very promising to develop a new tool in order to improve the forecasting of CO concentrations using weather forecasting.

ACKNOWLEDGEMENTS

This work is part of the coordinated research projects TIN2014-58516-C2-1-R and TIN2014-58516-C2-2-R supported by MICIN (Ministerio de Economía y Competitividad-Spain). Also, this work has been performed with the support of the Environmental Agency of the Andalusian Government, which provided the authors with all the monitoring data.

REFERENCES

[1] United States Environmental Protection Agency, What Are the Six Common Air Pollutants? Available online, http://www.epa.gov/air/urbanair/. Accessed on: 1 Apr. 2017.

[2] European Commission, Environment. Air Quality Standards. Available online, http://ec.europa.eu/environment/air/quality/standards.htm. Accessed on: 1 Apr. 2017.

[3] Fenger, J., Urban air quality. *Atmospheric Environment*, **33**, pp. 4877–4900, 1999.

[4] Vardoulakis, S., Gonzalez-Flesca, N. & Fisher, B.E.A., Assessment of traffic-related air pollution in two street canyons in Paris. *Atmos Environment*, **36**, pp. 1025–1039, 2002.

[5] Colbeck, I., *Environmental Chemistry of Aerosol*, Wiley-Blackwell: New York, 2008.

[6] Turias, I.J., et al., A competitive neural network approach for meteorological situation clustering. *Atmospheric Environment*, **40**, pp. 532–541, 2006.

[7] Turias, I.J., González, F.J., Martín, M.L. & Galindo, P.L., Prediction models of CO, SPM and SO2 concentrations in the Campo de Gibraltar Region, Spain: A multiple comparison strategy. *Environmental Monitoring and Assessment*, **143**, pp. 131–146, 2008.

[8] Martín, M.L., Turias, I.J., González, F.J., Galindo, P.L., Puntonet, C.G. & Gorriz, J.M., Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks. *Chemosphere*, **70**(7), pp. 1190–1195, 2008.

[9] Muñoz, E., Martín, M.L., Jiménez-Come, M.J., Turias, I.J. & Trujillo, F., Prediction of PM10 and SO2 exceedances to control air pollution in the Bay of Algeciras, Spain. *Stochastic Environmental Research and Risk Assessment*, **28**, pp. 1409–1420, 2014.

[10] Perez, P., Prediction of sulphur dioxide concentrations at a site near downtown Santiago, Chile. *Atmos Environ*, **35**, pp. 4929–4935, 2001.

[11] Brunelli, U., Piazza, V., Pignato, L., Sorbello, F. & Vitabile, S., Two days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. *Atmos Environ*, **41**, pp. 2967–2995, 2007.

[12] Comrie, A. & Diem, J.E., Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. *Atmospheric Environment*, **33**, pp. 5023–5036, 1999.

[13] Maffeis, G., Prediction of carbon monoxide acute air pollution episodes. Model formulation and first application in Lombardy. *Atmospheric Environment*, **33**, pp. 3859–3872, 1999.

[14] Pizarro, J., Guerrero, E. & Galindo, P., Multiple comparison procedures applied to model selection. *Neurocomputing*, **48**, pp. 155–173, 2002.

[15] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P., *Numerical Recipes in C*, 3rd ed., Cambridge University Press: Cambridge, 2007.