PREDICTION OF HOURLY OZONE CONCENTRATIONS WITH MULTIPLE REGRESSION AND MULTILAYER PERCEPTRON MODELS

C. CAPILLA
Polytechnic University of Valencia, Spain.

ABSTRACT

In this work ozone observations of an urban area of the east coast of the Iberian Peninsula, are analyzed. The data set contains measurements from five automatic air pollution monitoring stations (background suburban or traffic urban). The application of multiple linear regression and neural networks models is considered. These models forecast hourly ozone levels for short-term prediction intervals (1, 8, and 24 h in advance). The study period is 2010–2012. The input variables are meteorological observations, ozone and nitrogen oxides concentrations, and daily and weekly seasonal cycles. The performance criteria to evaluate the computations accuracy are the residual mean square error, the mean absolute error, and the correlation coefficient between observations and predictions. These criteria have better results for the 1-h and 24-h predictions in all the locations. The comparison of multiple linear regressions and multilayer perceptron networks indicates that the second approach allows to obtain more accurate forecast for the three prediction intervals.

Keywords: multilayer perceptron networks, multiple linear regression, ozone, urban air quality.

1 INTRODUCTION

The present study analyzes historical ozone (O_3) observations, with the aim of comparing models to predict in advance this pollutant's levels. Tropospheric O_3 is a secondary contaminant produced by the interaction of meteorological conditions, nitrogen oxides (NO_x) and volatile organic compounds. Its atmospheric dynamics is related to photochemical smog, acid rain, and climate change [1]. Information and alert hourly thresholds for O_3 have been established by the European Union Directive [2]. This directive gives monitoring and public information guidelines for air quality management systems. It also indicates objective values for the protection of vegetation and human health. Legislation on air quality improvement in Spain can be found in [3]. O_3 adverse effects on human beings have been reported by [4, 5]. Its strong oxidant properties, and interaction with other atmospheric pollutants and co-factors, affect the respiratory tract structure and function, especially amongst sensitive risk groups (children, elderly people, and individuals with respiratory illnesses).

The legislation implies mandatory public warning in case of thresholds exceedance. This has lead to a demand increase of short-term forecasting methods. Urban air quality managers have to monitor pollutants, predict their critical levels, communicate risk situations to public and authorities, and evaluate emission reduction strategies [6]. The application of statistical methods to analyze pollutants temporal variations has been largely considered. Raheem *et al.* [7] used a combination of multivariate statistical tools, to evaluate seasonal influences on O_3 concentrations in two Nigerian cities, and O_3 links with meteorological parameters and anthropogenic activities.

WIT CONFERENCES

This paper is part of the Proceedings of the 24th International Conference on Modelling, Monitoring and Management of Air Pollution (Air Pollution 2016)
www.witconferences.com

ISSN: 1743-7601 (paper format), ISSN: 1743-761X (online), http://www.witpress.com/journals
DOI: 10.2495/SDP-V11-N4-558-565

According to Ripley [8]: ' …Increasingly neural networks are being proposed to compute standard statistical procedures', and 'In one sense neural networks are little more than non-linear regression and allied optimization methods'. Gardner and Dorling [9] reviewed artificial neural networks applications (in particular the multilayer perceptron MLP approach) in the atmospheric sciences. They commented that if the relationship between input and output 'is non-linear then linear regression is clearly and inappropriate tool, although it may be possible to apply linear regression on a more local basis where the non-linearity can be dismissed'. Gardner [10] described the MLP method: 'All forms of linear regression, with and without data transformations and including arbitrary interactions between variables, can be considered as special cases of MLP neural networks', and 'As the temporal resolution increases from the daily to hourly timescale, non-linearities and more complex interactions are required in the models'. These non-linearities have not been completely 'appreciated', although they have been 'recognized'.

Elkamel *et al*. [11] studied a neural network approach to measure and predict O_3 levels around a heavily industrialized area of Kuwait. The neural network method gave superior predictions when compared with linear and non-linear regression models. Agirre-Basurko *et al*. [12] developed a MLR model and two multilayer perceptron (MLP) networks for O_3 forecasting, with several time intervals in the Bilbao urban area (Spain). They concluded that MLP had improved performance over MLR models.

The aim of this work is to compare the performance of MLR and MLP models, to predict hourly O_3 levels for short-term intervals (1 h, 8 h, and 24 h). The study area is in Valencia (Eastern coast of Spain). A background suburban and traffic urban monitoring stations are considered. The R [13], Statgraphics, and MATLAB (*Neural Network Toolbox*) programs are used. Previous work with data from a town close to this city [14], focused on the input variables to predict O_3 with MLP models. They studied different time windows and obtained daily predictions with accuracy and robustness. Castell-Balaguer *et al*. [15] applied descriptive methods to analyze the seasonal variations of O_3 in the Turia river basin, including two of these urban sites. They concluded that the average daily and monthly dynamics had summer/winter differences, depending on the location.

2 STUDY AREA AND DATA SET

Five monitoring stations are considered. They are located in the urban area of Valencia (Spain). Table 1 gives the characteristics of these sites. The Mediterranean Centre for Environmental Studies Foundation applied quality control methods to the data set used in this work [16]. The report [17] assessed air pollution health impacts in Valencia. In their study period, air pollutants emissions were a consequence of motor vehicles. Their results showed that an increase of 1.3% in the daily deaths number, was associated with a 10% increase in O_3 concentration. This percentage was also linked to a 1.1% increase in circulatory diseases admissions, and a 6.1% of EPOC and 6.3% asthma emergencies.

The air quality data used in the analysis are hourly average O_3 and NO_x concentrations. The study period is 2010–2012. Information and alert hourly thresholds for O_3, are 180 and 240 μgr/m^3. These limits, established for the protection of human health, were not exceeded in any of the stations during 2012 [18]. The meteorological observations are wind speed (WS), wind direction (WD), temperature (T), relative humidity (RH), pressure (P), and solar radiation (SR). These parameters are observed in Pista de Silla station, and all these data are also used in the analysis of the other stations, except in Avd. Francia where WS and WD measurements are also registered. Table 2 shows the descriptive analysis (mean and standard errors) of the air pollution and climatic variables. Average O_3 level has been higher in 2012

Table 1: Monitoring sites of the urban area of Valencia.

Name	Type	Variables	Coordinates	Altitude (m)
Avd.Francia	Traffic urban	Air quality meteorological	0°20'34''W 39°27'29''N	7
Molí del Sol	Traffic urban	Air quality	0°24'30''W 39°28'52''N	15
Pista de Silla	Traffic urban	Air quality meteorological	0°22'36''W 39°27'29''N	11
Politècnic	Traffic urban	Air quality	0°20'15''W 39°28'47''N	7
Viveros	Background suburban	Air quality	0°22'10''W 39°28'46''N	11

Table 2: Means and standard errors of the variables used in the analysis.

	2010		2011		2012	
Station	Mean	Standard error	Mean	Standard error	Mean	Standard error
Avd.Francia						
O_3	45,84	25,55	45,05	27,49	50,2	27,44
NO_x	51,57	57,69	52,12	59,08	39,54	45,11
WS	1,3	1,3	1,3	1,3	1,3	1,3
WD	169	101,8	161	103	167	100,3
Molí del sol						
O_3	48,64	32,68	47,54	32,51	54,18	35,03
NO_x	60,59	63,64	51,8	52,22	48,68	51,28
Pista Silla						
O_3	44,46	30,15	47,93	31,59	38,39	24,51
NO_x	87,65	78,08	64,67	59,39	74,31	76,58
WS	0,9	0,8	0,7	0,7	0,6	0,7
WD	207	108,8	184,6	112,3	200,9	107,8
T	17,8	6,9	18,8	6,39	17,5	6,85
RH	62	15,8	65,46	14,69	61,5	19,72
P	1009,9	6,8	1013,5	6,44	1015,6	7,47
SR	138,8	231,4	150	237,8	161,5	261,6
Politècnic						
O_3	51,89	29,37	53,62	31,64	54,05	31,78
NO_x	51,58	53,84	46,48	53,19	28,22	29,43
Viveros						
O_3	43,52	26,38	43,35	29,81	39,34	29,64
NO_x	46,69	40,27	45,70	49,46	53,67	58,74

Table 3: Number of observations available with complete records.

Station	Training set	Validation set	Test set
Avd.Francia			
1 h	7579	7701	5691
8 h	7560	7672	5670
24 h	7543	7642	5642
Molí del Sol			
1 h	7001	7420	6709
8 h	6966	7386	6690
24 h	6932	7363	6675
Pista de Silla			
1 h	7055	7974	6747
8 h	7022	7915	6709
24 h	6993	7845	6692
Politècnic			
1 h	8150	7233	7325
8 h	8079	7170	7278
24 h	8027	7107	7243
Viveros			
1 h	8054	6845	3537
8 h	8038	6810	3474
24 h	8065	6767	3352

in Avd. Francia, Molí del Sol, and Politècnic. In the other two stations, the lowest mean value of this pollutant was observed in 2012. The average hourly O_3 and NO_x levels of 2010 and 2011 were smaller in the background suburban site (Viveros) than in the other sites. In 2012, the smallest mean of O_3 hourly values is in Pista Silla, although this station has the highest NO_x hourly average.

The prediction models are MLR and MLP networks. The input variables, following the results of [14], are meteorological observations and pollutants concentrations. Daily and weekly seasonalities are included as predictors, in the form of sine and cosine components. These seasonalities were descriptively analyzed by [15], in Pista Silla and Viveros. Hourly forecasts were obtained with 1, 8, and 24 h in advance. The number of neurons in the hidden layer of the MLP networks, was from 5 to 55. The hyperbolic tangent transfer function and the Levenberg–Marquard algorithm were applied. The observations were separated into the training (year 2010), validation (year 2011) and, test (year 2012) sets. The MLR models were estimated with the observations of 2010, and their evaluations were done with the predictions for 2012. Table 3 gives the number of hourly observations used in analysis, for the three prediction intervals.

Three parameters were computed to evaluate the models performance: the root mean square error (RMSE), the mean absolute error (MAE), and the correlation coefficient between observations and predictions (r).

3 RESULTS AND DISCUSSION

The models evaluations results are presented in Tables 4 and 5. Table 4 contains the MLR performance criteria values, for the five monitoring sites and the three prediction intervals. These criteria values are better for Viveros site when the prediction interval is 1 h. If the

Table 4: RMSE, MAE and r values for the MLR predictions.

Station	RMSE	MAE	R
Avd.Francia			
1 h	9,6688	6,9949	0,9373
8 h	21,2017	17,1604	0,6582
24 h	17,9938	14,2254	0,7679
Molí del Sol			
1 h	10,339	7,6434	0,9553
8 h	25,5284	20,8837	0,7045
24 h	20,6767	16,1778	0,8125
Pista de Silla			
1 h	9,4452	7,0286	0,9237
8 h	20,8772	16,7847	0,5371
24 h	18,2029	14,5048	0,6746
Politècnic			
1 h	9,9324	7,2491	0,9516
8 h	23,1183	18,8079	0,6169
24 h	18.9615	14,9854	0,8131
Viveros			
1 h	9,2395	6,7192	0,9511
8 h	22,6609	18,4089	0,6492
24 h	18,8583	15,1749	0,7009

Table 5: N_h, RMSE, MAE, and r values of the MLP predictions.

Station	N_h	RMSE	MAE	r
Avd.Francia				
1 h	10	8,9024	6,4417	0,9478
8 h	10	18,5894	14,5120	0,7579
24 h	5	17,8682	14,1471	0,7692
Molí del Sol				
1 h	10	9,3206	6,7982	0,9651
8 h	35	22,2441	17,5646	0,7865
24 h	20	20,7372	16,2411	0,8064
Pista de Silla				
1 h	5	8,4694	6,2789	0,9393
8 h	15	19,0643	15,0111	0,6674
24 h	5	18,1307	14,3614	0,6811
Politècnic				
1 h	15	9,1747	6,7397	0,9608
8 h	10	21,2384	16,7431	0,7762
24 h	20	19,1707	15,1257	0,8084
Viveros				
1 h	15	8,4704	6,0507	0,9605
8 h	25	20,9776	16,0970	0,7247
24 h	10	19,7404	15,0829	0,7549

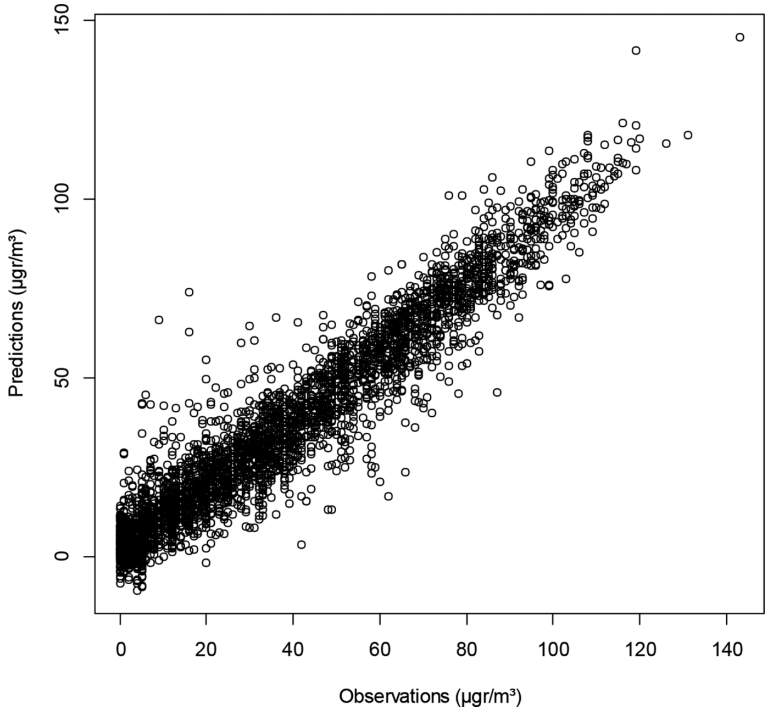

Figure 1: Observations and predictions at Viveros site 1 h in advance.

forecasts are computed 8 h in advance, the MLR has better evaluation results for Pista Silla station in terms of RMSE and MAE. Nevertheless, the linear correlations between observations and predictions are higher in the other sites. Predictions RMSE and MAE are better in Avd.Francia, for 24 h ahead calculations. However the correlation coefficient r is higher in Molí del Sol and Politècnic.

The MLP networks evaluation is shown in Table 5. It contains the best neural network result, with its number of hidden units (N_h). Pista Silla and Viveros site have similar RMSE and MSE results for the 1 h ahead prediction. The best correlation coefficient values are in Molí del Sol and Politècnic locations. The forecast 8 h in advance is better in Avd.Francia with lower RMSE and MAE values, but in Molí del Sol and Politècnic the correlation coefficient is higher. The same results are observed in case of the 24-h interval. In all stations the best predictions with MLR and MLP models are obtained for the 1 h interval, and the worse ones for the 24-h interval. Comparisons of the two approaches, indicates that MLP allows to predict with more accuracy than MLR. Observations and 1 h ahead predictions for Viveros, are represented in Fig. 1.

4 CONCLUSIONS

Multiple linear regression and multilayer perceptron models have been used to predict ozone with three intervals. In the five monitoring sites of the study, the second method gave better results. The prediction for the shortest lag was more accurate. These models were applied in [12] to forecast ozone concentrations in Bilbao (Spain). In the four stations of this urban area, the MLR had worse correlation coefficients than the MLP, with traffic, nitrogen dioxide

concentrations, meteorological parameters, and seasonal cycles as inputs, when the prediction intervals were 1 or 8 h. These coefficients were smaller than the ones in Tables 4 and 5. Gómez-Sanchis *et al*. [14] obtained MAE results between 5,87 and 7,21, and RMSE values from 7,31 to 9,04, when predicting one day ahead O_3 concentrations. They used as inputs daily O_3, nitric oxide, nitrogen dioxide, and climatic variables in a MLP model, 'applied to a small town near Valencia'.

REFERENCES

[1] Hartman, D.L., Kleintank, A.M.G., Rusticucci, M., Alexander, L.V., Brönnimann, S., Charabi, Y., Dentener, F.J., Dlugokencky, E.J., Easterling, D.R., Kaplan, A., Soden, B.J., Thorne, P.W., Wild, M. & Zhai, P.M., Observations: atmosphere and surface (Chapter 2). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds. T.F. Stocker, D. Qin, G.–K. Plattner, M. Tigno, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P.M. Midgley, Cambridge University Press, Cambridge: United Kingdom and New York, NY, USA, pp. 159–218, available at http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter02_FINAL.pdf (accessed 25 January 2016), 2013.

[2] Official Journal of the European Union, *Directive 2008/50/EC of the European Parliament and of the Council of 21 May2008 on Ambient Air Quality and Cleaner Air for Europe,* available at http:// eur-lex.europa.eu/legal-content/AUTO/?uri=CELEX:32008L0050&qid=145371475043&rid=1 (accessed 25 January 2016).

[3] Boletin Oficial del Estado. *Real Decreto102/2011 de 28 de Enero, Relativo a la Mejora de la Calidad del Aire,* available at http//www.boe.es/boe/dias/2011/01/29/pdfs/BOE-A-2011-1643.pdf (accessed 25 January 2016).

[4] World Health Organization, *Effect of Air Pollution on Childrens Health and Development.* World Health Organization Regional Office for Europe, Copenhagen, Denmark, 2005, available at http://apps.who.int/iris/bitstream/10665/107652/1/E86575.pdf (accessed 25 January 2016).

[5] World Health Organization, *Air Quality Guidelines. Global Update 2005. Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide,* World health Organization Regional Office for Europe, Copenhagen, Denmark, 2006, available at http://www.euro.who.int/__data/assets/pdf_file/0008/147851/E87950.pdf (accessed 25 January 2016).

[6] Dutot, A.L., Rynkiewicz, J., Steiner, F.E. & Rude, J., A 24-hour forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling & Software*, **22**, pp. 1261–1269, 2007. http://dx.doi.org/10.1016/j.envsoft.2006.08.002

[7] Raheem Abdul, A.M.O., Adekola, F.A. & Obioh, I.O., The seasonal variation of ozone, sulphur dioxide and nitrogen oxides in two Nigerian cities. *Environmental Modeling & Assessment,* **14**, pp. 497–509, 2009. http://dx.doi.org/10.1007/s10666-008-9142-x

[8] Ripley, B.D., Statistical aspects of neural networks (Chapter 2). *Networks and Chaos-Statistical and Probabilistic Aspects,* eds. D.E. Barndorff-Nielsen, J.L. Jensen & W.S. Kendall, Chapman & Hall/CRC: London, UK, pp. 40–123,1993.

[9] Gardner, M.W. & Dorling, S.R., Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. *Atmospheric Environment,* **32**(4), pp. 2627–2636, 1998. http://dx.doi.org/10.1016/S1352-2310(97)00447-0

[10] Gardner, M.W., The advantages of artificial neural networks and regression tree based air quality models (Chapters 1 and 2). *A dissertation submitted to the School of Environmental Sciences of the University of East Anglia* (part of the requirements for the degree of Doctor of Philosophy), UK, 1999.

[11] Elkamel, A., Abdul-Wahab, S., Bouhamra, W. & Alper. E., Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach. *Advances in Environmental Research,* 5, pp. 47-59, 2001. http://dx.doi.org/10.1016/S1093-0191(00)00042-3

[12] Agirre-Basurko, E., Ibarra-Berastegui, G. & Madariaga, I., Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. *Environmental Modelling & Software*, **21**, pp.430-446, 2006. http://dx.doi.org/10.1016/j.envsoft.2004.07.008

[13] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, available at available at http://www.R-project.org/, 2014.

[14] Gómez-Sanchis, J., Martín-Guerrero, J.D., Soria-Olivas, E., Vila-Francés, J., Carrasco, J.L. & del Valle-Tascón, S., Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. *Atmospheric Environment,* **40**, pp. 6173-6180, 2006. http://dx.doi.org/10.1016/j.atmosenv.2006.04.067

[15] Castell-Balaguer, N., Téllez, L. & Mantilla, E., Daily, seasonal and monthly variations in ozone levels recorded at the Turia river basin in Valencia (Eastern Spain). *Environmental Science & Pollution Research*, **19**,pp. 3461-3480, 2012.

[16] Castell-Balaguer, N., Téllez, L., Luján, A. & Mantilla, E., *Informe Final Previozono 2010. Programa Especial de Vigilancia de las Concentraciones de Ozono Troposférico en la Comunidad Valenciana.* http://www.agricultura.gva.es/documents/20549779/161530536/informe10.pd (accessed 28 January 2016), 2010. http://dx.doi.org/10.1007/s11356-012-0881-5

[17] Ballester, F., Iñíguez, C. & García, F., *ENHIS-1 project: WP5 health impact assessment. Local city report Valencia*. http://www.apheis.org/CityReports2005/Valencia.pdf (accessed 4th May 2006), 2005.

[18] The Mediterranean Centre for Environmental Studies Foundation, *Informe Final Previozono 2012,* available at http://www.agricultura.gva.es/documents/20549779/161530540/informe12.pdf (accessed 28 January 2016), 2012.