METHODOLOGICAL CONSIDERATION ON PRE-PROCESSING DATA OPTIMIZATION CONCERNING AIR DISPERSION MODEL AND NEURAL NETWORKS: A CASE-STUDY OF OZONE PREDICTION LEVEL

A. Pelliccioni, R. Cotroneo and F. Pungì

ISPESL-DIPIA, Via Fontana Candida 1, 00040, Monteporzio Catone (RM), Italy

ABSTRACT: This work analyzes the results of a Neural Network model applied to air pollution data. In particular, we forecast ozone pollutants levels in a short term using both air dispersion models and neural network methods. The purpose of this work is to provide a novel methodological procedure to analyze environmental data by using a neural net as forecast technique for ozone levels in the urban area of Rome. Results show that the model performance can be improved by pre-processing input data using typical data-mining techniques and coupling air dispersion model with neural net.

Key words: *Ozone, neural networks, Data mining, Stepwise algorithm selection, resampling*

1. INTRODUCTION

Air quality problems, produced by high levels of ozone, have an impact on human health related with respiratory diseases. The latter are critical especially in large metropolitan areas where there is relevant exposure of population with consequent health problems. Ozone is a reactive gas and its levels are strongly dependent both from regional transport and from the micro-meteorological conditions of the site and the seasonal effects. The prediction of Ozone levels is very complex to obtain as described in different studies (Gardner et al., 2000; Dutot et al., 2007). For Ozone models one of the most difficult problems to deal with is the simulation of the chemical reactions that occur in atmosphere, linked to the long range transport, to the incoming solar radiation and to the atmospheric turbulence conditions (Penket et al., 2004).

Among the complex systems, an important methodology to forecast air pollution data by advanced statistical methods are the neural networks (NNs), that can work as *universal approximators* of non-linear functions and, consequently, can be used in assessing the dynamics of such systems. NN methods have been developed for forecasting daily maximum ozone levels in various urban areas, using average daily meteorological data as input parameters (Comrie, 1997). All above environmental applications use the NN model as regression tool.

Our approach consists of a combination of statistical models (neural nets) and deterministic models (air dispersion models), to improve the accuracy of the ozone predictions. In fact, in our work we developed an integrated modelling system coupling an air dispersion model (named as FARM (Gariazzo et al., 2007) with a neural network method in order to adjust the influence of important variables on air dispersion models and, contemporaneously, to minimize the number of input variables to the NN model. Further, we provided a methodological procedure in order to optimize pre-processing inputs (both variables and patterns), using the data mining (DM) techniques, to achieve the best performance for the data. For environmental data, the optimization of the input patterns is a strategic point because from the modelling point of view the simulation of extreme events (high values of air pollution) has a legislative concern.

2. DATASET DESCRIPTION

The city of Rome is characterized by high concentrations of ozone, associated with hot sunny days and stagnant conditions. In the urban area of Rome, we analysed three air pollution episodes: *Episode A*: 20-24 June 2005 (photochemical pollution + intensive campaign); *Episode B:* 25-29 July 2005 (photochemical pollution); *Episode C*: 9-13 January 2006 (winter pollution).

The dataset includes about 16 variables and 1200 patterns collected from 5 monitoring stations. The temporal accounting for each pattern is hourly, corresponding to the time data collections. The 16 variables concern the following three groups:

- Pollution variables observed by monitoring stations: Carbon monoxide (CO in mgm^{-3}); Nitrogen Oxide (NO2 in μgm^{-3});Ozone (O3 in μgm^{-3})
- Meteorological variables: Temperature 32 and 10 meter (C°); Atmospheric boundary layer height (m) [H_{MIX}]; Monin-Obukhov length (m^{-1}), [1/L] Wind speed and velocity at 32 and 10 meter [WS_{10m} and WS_{32m}]; Global Solar Radiation (Wm^{-2}) [GSR]; relative humidity [UR] (%); Net Solar Radiation (Wm^{-2}) [NSR]; Pressure (mbar) [P]. The meteorological field, campaign conducted from June 2005 to June 2006, represent three typical meteorological conditions of the studied area
- Pollutants calculated by an air dispersion model –FARM-: Carbon Monoxide (CO_{FARM}); Nitrogen Oxide ($NO2_{FARM}$); Ozone ($O3_{FARM}$)

3. METHODOLOGY AND DATA PRE-PROCESSING

In this work, the dataset has been analysed using two type of Artificial Neural Network model with a supervised algorithm: the Reference Neural Network (RNN) (Han and Kamber, 2001) applied to the discretised O3 variable and the Forecast Neural Network (FNN) applied to the continuous O3variable. For a Neural Network, each pattern is composed by a selected choice of both air pollution and meteorological variables and ozone levels (target variables). The data have been analysed starting from two separated phases (see figure 1) in which we utilised:

- meteorological and monitoring air pollution data and other primary pollutants by air dispersion model (DS(FARM) related to variables space and $DS(FARM)^{RESA}$ related to units space pre-processing) to calculate ozone
- meteorological and monitoring air pollution data by monitoring stations (DS related to variables space and DS^{RESA} related to units space pre-processing) to calculate ozone

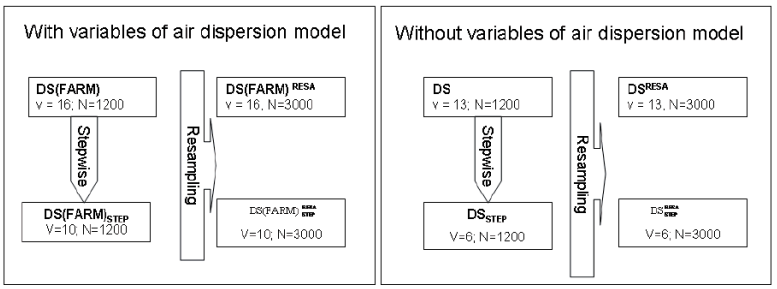

Figure 1. Dataset scenarios.

Before to proceed with the Neural Network application, we need to apply a data pre-processing methodology to simplify the dataset by reducing multidimensional datasets to lower dimensions for analysis. We applied two conventional DM filters of data pre-processing: the stepwise algorithm to select the best variables ($DS(FARM)_{STEP}$ and DS_{STEP}- see figure 1) according to the well know "parsimony principle", and the resampling technique ($DS(FARM)^{RESA}$, $DS(FARM)^{RESA}_{STEP}$, DS^{RESA} and DS^{RESA}_{STEP}) to improve the weights of patterns related with high levels – outliers - of ozone (greater than 80 μgm^{-3}). At the end of data pre-processing, both in the variables space and in the patterns space, it was applied the Reference Neural Network (RNN) that was used to provide a measure (tendency) of level of learning of the net. It also provides a value of generalization and adaptation of the model to the data. At the end of above processing, we use advanced Forecast Neural Network (FNN), having as target the best performance model. The FNN should predict the target variable in a correct and more accurate way respect to the RNN.

4. VARIABLES PRE-SELECTION

The NN robustness is determined introducing significant input variables and removing, if it is possible, the least important ones. In addition, using all variables of dataset could increase the complexity of the training model and produce the well known overfitting. For this reason, it has been selected a restricted variables subset (decreasing the dimension of the variables space) to feed the NN, trying to obtain a good fitting of the model to the data without also increasing its complexity and without introducing correlation among the input variables (multicolinearity).Such selection was carried out by the classic statistics technique of stepwise algorithm (Jobson, 1992), that examines variables more significant with the highest partial F value. This algorithm was applied to DS(FARM) and DS.

Variables	DS(FARM)	$DS(FARM)_{STEP}$	DS	DS_{STEP}
Station	X	Rank 7	X	Rank 10
CO_{FARM}	X	Rank 4		
$NO2_{FARM}$	X	Rank 9		
$O3_{FARM}$	X	Rank 1		
CO	X	Rank 6	X	Rank 6
NO2	X	Rank 3	X	Rank 2
$T_{10\,m}$	X		X	Rank 3
$T_{32\,m}$, WS_{10m},	X		X	
H_{MIX}	X	Rank 2	X	Rank 1
1/L	X	Rank 10	X	Rank 7
WS_{32m}	X		X	Rank 4
GSR	X		X	Rank 8
RH	X	Rank 5	X	Rank 5
NSR	X		X	Rank 9
P	X	Rank 8	X	Rank 4
Total	16	10	13	10

The results were synthesized in the Table 1. For each dataset, we utilised the variables with the highest partial F value, which in Table 1 are indicated with rank (rank 1 for the highest value, then rank2 …).

It can be observed that the inclusion of variables calculated by FARM in the dataset as inputs of NN allows to keep out some conventional variables (as temperature, wind speed direction, solar radiation, etc.) that usually are used into classical application of neural networks. In this way, the air dispersion model variables used as NN inputs synthesize the effects of turbulence on the pollutants levels and they could be better respect to the direct variables to train the NN. Generally, it is useful to modelling the pollutants levels with a minimum number of variables. In this work we applied the parsimony principle (namely also Ockham's Razor) exploring those selections that minimize the input variables without losing information contained in the dataset. We decided to work with an easier dataset to verify the accuracy and the effectiveness of the neural network in presence of the minimal number of significant variables.

5. PATTERNS PRE-SELECTION

Patterns selection used for neural network is one of the most important tasks that should be solved in order to achieve good generalising capabilities of the model. In our case, Ozone's distribution is highly skew. In fact, the 71% of units belonging in the first class (0-20 μgm^{-3}), and in these extremes we have only ten values (1%). For this reason, the RNN is not able to train the extreme classes. Therefore, we decided to proceed applying resampling techniques (computer-intensive methods) to neural network validation in presence of outliers. The resampling method is able to obtain the benefits of statistics and forecasting. It permits, using only the available observed data, to obtain a very high number of samples starting from the same population (Witten and Frank, 2005). So far, using the pseudos-samples it becomes easier to make learning the outliers to the neural network, considering that every class is represented with the same proportion of the initial dataset (Bradley, 1997). A sampling strategy is given as a guide for investigations, when there is no previous knowledge about the structure of the patterns. We proceeded to patterns resample (250% with a bias to Uniform class=0.2). In this way, we were able to make training the Neural Network also the outliers (see Fig. 2).

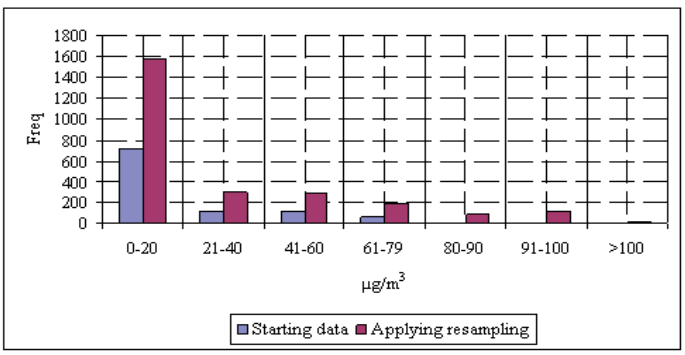

Figure 2. Ozone distribution ($\mu g/m^3$).

Therefore, it has been possible to apply a NN able to learning the extreme events. It is emphasized that with the resampling was possible to make learning the RNN also the data of the extreme classes that couldn't be learned in the previous situation (DS(FARM); DS). In particular, we observed that the variables reduction related to $DS(FARM)^{RESA}_{STEP}$ and DS^{RESA}_{STEP} had no effects both on the Neural Network performance and on learning of the outliers.

6. COMPARED RESULTS ABOUT REFERENCE NEURAL NETWORK AND FORECAST NEURAL NETWORK

All above pre-processing results have been obtained using as NN a RNN. Now, we use FNN to develop and optimise the NN, obtaining the best model Multilayer Perceptron (M.L.P.) configuration, using the main suggestions coming from pre-processing results.

In our simulations, we tested the FNN performances by using different percent of input data during the training phase. In particular, during the training we considered all input data (100%) and 50% of the selected patterns. Having chosen 50% of input data during the training phase, the remaining 50% has been used for the generalization phase and to evaluate the accuracy of the model. The results are always referred to generalisation phase and they are calculated using all data (100%) and half dataset (50%) respectively for the training and selection phases. During the training phase, we use the conventional sigmoidal activation functions (see Eq. 1).

$$F(P) = \frac{1}{1 + e^{-(p-s)}} \tag{1}$$

Different types of selections of NN algorithm to perform the weights corrections can be used during the training. The popular backpropagation algorithm was used to train our data (using as learning rate two values 0.1 and 0.05). Another important parameter to the best optimization of FNN, concerns the number of hidden layer neurons. We

attempted different selection of hidden neurons number, testing 8, 10 and 12 neurons. At the end of the simulation phase, the best result was obtained with 12 neurons. It will be used the well known square correlation coefficient (namely also coefficient of determination, R^2) as a statistical measure of the global fit of the model. In table 3 are shown the results of the generalisation phase in the case of dataset with air pollutants both in case of stepwise algorithm and resampling. As datasets simulations we consider those described in figure 1. It was given the R^2 related to training (R^2 Train) and selection (R^2 Selection) phases and to the total amount of data (R^2 Total). It can be observed that the R^2 is always very good in both phases (>0.90), showing an excellent forecast for the ozone levels by FNN. Further, the model is able to reproduce data never used during the training phases (R^2 selection) in good agreement with the experimental ones. In fact, during the training phase, in every dataset it was obtained R^2>0.95, whereas for the generalization a value of R^2>0.94 was achieved.

Table 1. Synthesis of results - with air pollutes (CO, NO_2).

Phase		$DS(FARM)_{STEP}$			DS_{STEP}			$DS(FARM)^{RESA}_{STEP}$			DS^{RESA}_{STEP}		
Train(%)	Selection(%)	R^2			R^2			R^2			R^2		
		Total	Train	Selection	Total	Train	Selection	Total	Train	Selection	Total	Train	Selection
100	0	0,97	0,97		0,97	0,97		0,97	0,97		0,99	0,99	
50	50	0,94	0,95	0,94	0,97	0,98	0,96	0,98	0,98	0,97	0,99	0,99	0,98

Other important considerations can be showed by the above table taking into considerations the selection data results. Our simulations show that applying also the resampling techniques we have better results than applying stepwise only.

Our study suggests further considerations for short term predictions of air pollution levels by neural networks models. Usually, high pollutants levels (e.g. ozone outliers in our simulations) are forecasted by means of NN without making any rational choice respect to the input variables and the patterns selection's techniques. The purpose of this work was to select inputs to optimize the information inside the main environmental variables (such as meteorologicals, monitoring pollutants, source positions, etc.) to make a model with MLP that is a *universal approximator*. The air dispersion model variables used as NN inputs is a novelty in the environmental model. In a previous work [10] many studies were conducted in applying this coupling methodology, but data mining techniques were never used as data pre-processing.

These results show that NN provides the same performance respect to those obtained using only observed concentrations as NN inputs. This justifies the use of air dispersion model variables as input to the NN (e.g. calculated by the FARM model) and a few other variables by monitoring stations..
Another relevant result obtained by our study concerns the simulations of extreme events, such as high ozone levels,. Whereas conventional approaches are not able to forecast these events (mainly because of they are rare and consequently during the training the NN cannot simulate them), the resampling techniques applied during the training phase, succeed in taking into account these situations.

The following Figures 3 and 4 show O_3 predicted by NN with air dispersion model variables and without air dispersion model variables. After resampling, you can also observe that the NN learns better our input data in the extreme class . In fact, resampling technique gives excellent results in estimating the future performance of FNN (R^2=0.95 and R^2=0.97).

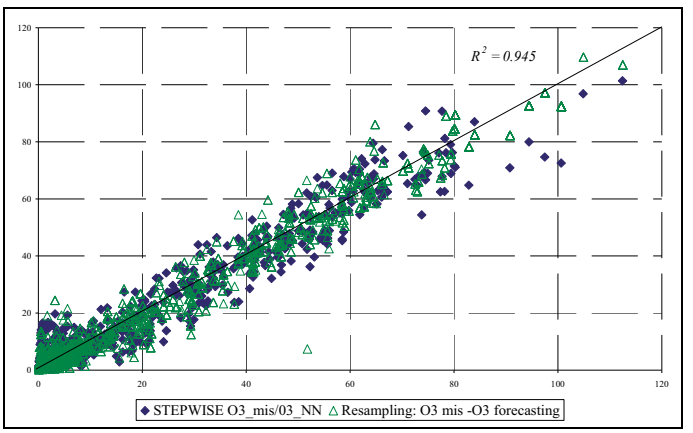

Figure 3. O_3 predicted by $DS(FARM)_{STEP}$ and $DS(FARM)^{RESA}_{STEP}$.

In order to underline the importance of the outliers, in the environmental field, we consider the performance of the NN related to the four datasets and divide them in two classes of ozone levels: lesser (<80 μgm^{-3}) or greater (>80 μgm^{-3}) than 80 μgm^{-3}. Ozone levels greater than 80 μgm^{-3} have an impact on health and they are extremely rare. In

Table 2 the absolute value between measured ozone levels and predicted values by NN (|O3mis-O3NN|) related to the two classes defined above are given.

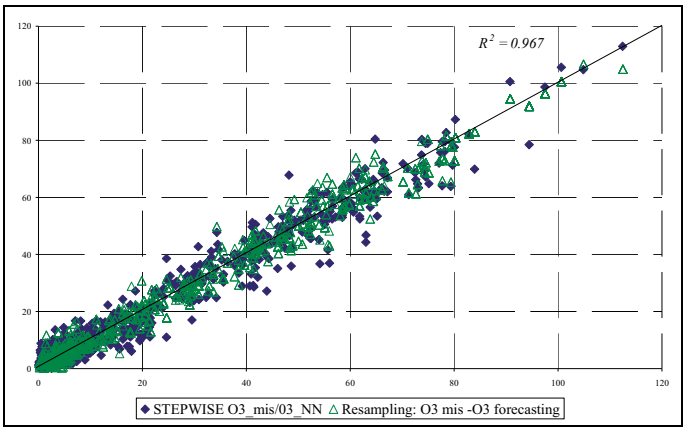

Figure 4. O_3 predicted by DS_{STEP} and DS^{RESA}_{STEP} .

The results show that the resampling techniques are a right choice to forecast the high pollutants ozone levels. In fact, while the error is similar for the lower class (average 3,15 μgm^{-3} vs. 2,85 μgm^{-3}), it is smaller in the higher class (average 9,99 μgm^{-3} vs. 3,36 μgm^{-3}).

Table 2. Outliers Results (mean of absolute distance) – only correct classified patterns.

	$DS(FARM)_{STEP}$		DS_{STEP}		$DS(FARM)^{RESA}_{STEP}$		DS^{RESA}_{STEP}	
O3	N	|O3mis-03NN|	N	|O3mis-03NN|	N	|O3 -03NN|	N	|O3 -03NN|
<80 (μgm^{-3})	894	3,55	899	2,75	2065	2,96	1447	2,74
>80 (μgm^{-3})	10	14,29	10	5,70	206	4,65	206	2,06
Total	904	3,67	909	2,78	2271	3,11	1653	2,66

6. CONCLUSION

The aim of our work consisted in the investigation of methodological considerations on pre-processing data to optimise the input information provided to a neural network model used to forecast air pollutant levels in a complex situations such as the urban area of Rome. We used as environmental data those coming from urban pollutants by monitoring stations, meteorological stations and concentrations calculated by an air dispersion model. We tested both the stepwise algorithm and the resampling technique as pre-processing data to obtain the best performance by a 3-layer perceptron model. Identification of key variables is important for enhancing knowledge of ozone prediction. Results seem to suggest that is better to proceed with a two steps analysis (first with an air dispersion model then with neural network) rather than make a choice involving all input variables (air pollution, meteorological variables, exc.) as in conventional NN applications. In this way, it is possible to adopt the air dispersion model variables as inputs of the NN using a few monitoring stations variables. The data reduction also allowed to improve the computational efficiency of the applied FNN maintaining the same performance obtained with the full dataset. Moreover, the resampling techniques provide a better description for extreme events of our FNN that in general don't succeed to forecast these events.

REFERENCES

Bradley, E., 1997: Bootstrap Methods: Another Look at the Jackknife. Stanford University.

Comrie R.S, 1997: Comparing neural network and regression models for ozone forecasting. *J. of the Air and Waste Management Association,* **47,** 653-663.

Dutot, A.L., Rynkiewicz, J., Steiner, F.E. and J.Rude, 2007: A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling & Software*, **22**, 1261-1269.

Gardner, M.W and S.R. Dorling, 2000: Statistical surface ozone models: an improved methodology to account for non-linear behaviour*. Atmospheric Environment, **34,*** 21-34.

Gariazzo, C. Silibello, C. Finardi, S..Radice, P Piersanti, A. Calori, G. Cecinatoc, A. Perrinoc, C.Nussio, F Cagnoli, M. Pelliccioni, A. Gobbie G. P. and P. Di Filippo, 2007: A gas/aerosol air pollutants study over the urban area of Rome using a comprehensive chemical transport model. *Atm. Env.,* **41,** 34, 7286-7303.

Han, J. and M. Kamber, 2001: Data mining: concepts and techniques. Morgan Kaufmann, San Francisco,.

Jobson, J. D., 1992: Applied Multivariate Data Analysis. *Categorical and Multivariate Methods*, **2.**

Pelliccioni, A. and T. Tirabassi, 2006: Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. *Environmental Modelling & Software,* **21**, 4, 539-546.

Penkett, S.A Evans, M. J. Reeves, C. E. .Law, K.S. Monks, P.S .Bauguitte, S. J.B Pyle, J. A. Green, T. J. Bandy, B. J. Mills, G. Cardenas, L. M. Barjat, H. Kley, D. Schmitgen, S. Kent, J. M. Dewey, K. and J. Methven, 2004: Long-range transport of ozone and related pollutants over the North Atlantic in spring and summer, *Atmos. Chem. Phys. Discuss.*, **4,** 4407-4454

Witten, H.I. and E. Frank, 2005: Data Mining Practical Machine Learning Tools and Techniques. Second Edition, Morgan Kaufmann.