Presented at the 94th AMS Annual Meeting, Atlanta, GA, Feb 2-6, 2014

1.3

OPTIMIZATION OF NEURAL NETWORK BY USING CLUSTER ANALYSIS FOR OZONE

A. Pelliccioni[1] and R. Cotroneo[2]

1. Inail- Research, Via Fontana Candida 1, 00040 Monteporzio Catone, Italy
2. Istat, Viale dell'Oceano Pacifico 171,00198 Roma Italy

1. Introduction.

As is well known, air quality problems, due to the ozone (O_3) could produce effects on human health related to respiratory problems, damage to ecosystems, agricultural crops and materials (World Health Organization –WHO-, 2003). The ozone is classified as a secondary pollutant and its levels are determined by complex photochemical reactions with primary pollutants (EPA, 2006). The concentrations are strongly dependent both on micro-meteorological conditions linked to turbulence and to the effects related to the seasons (Finlayson-Pitt, 1986), (Seinfeld, 1986). As described in different works, the prediction of ozone levels is very complex to obtain by mathematical models (Comrie, 1997), (Gardner, 2000), (Gardner, 1998), (Pelliccioni, 2010a,b).
In this work, NNs have been developed to forecast ozone levels in complex systems using hourly pollution and meteorological data as input parameters. In this work, in order to improve the NN ability to "capture" the true hidden relation inside the data set by means of the pre-selection of information, we applied the cluster analysis techniques, as preprocessing technique. In particular, we demonstrated that the consistency of the ozone forecast is dependent on the selection of patterns during the training phase, and the results obtained depend on the statistical distribution of data.
The aim of this paper is to investigate the usefulness of NN to predict ozone levels by using the information coming from a machine learning algorithms. Further, we presented a new strategy for the pattern optimisation to forecast ozone that leads to a better performance of NN.

2. Materials and methods

2.1. Description of used data set

The environmental data set come from the urban background monitoring station of the ARPAL (Environmental Protection Agency of Lazio Region) of Rome (Villa Ada monitoring station), and consists of hourly data during the two calendar years 2006 and 2007.
The Villa Ada station collects a set of meteorological variables such as solar radiation, temperature and humidity, and some conventional air pollutants variables (O3, NO, NO_2, CO).
Ours data set regards about 14324 hourly patterns, gathered every day. The variables used are the following:

- observed pollutant variables:
  - Carbon monoxide (mg/m^3)- CO
  - Nitrogen Oxide (µg/m^3) -NO
  - Nitrogen Dioxide (µg/m^3) -NO_2
  - Ozone (µg/m^3) - O_3

- meteorological variables:
  - Temperature (C°) - T
  - Global Solar Radiation (W/m^2) - GSR
  - Relative Humidity - RH (%)

Table 1 shows the general statistics calculated for 2006 and 2007.
The table examines the main statistical parameters (for each year): mean, standard deviation, maximum and minimum, variation coefficient (CV) that represents the ratio of the standard deviation to the mean.
Our dataset shows the maximum hourly value of ozone for 2006 is around 227.1µg/m³ and 189.1µg/m³ for 2007, verified during the summer season (15/07/2007 h.15.00). In 2007 the maximum hourly value of CO is about 4.1 mg/m³ during the winter season (11/01/2007 h.23.00) compared with the maximum hourly value of CO is about 3.8 mg/m³ in 2006.
For 2007, we also observed the maximum value of the variability in the series (CV=102.3%) for O_3.

[1] *Corresponding Author Address*: Armando Pelliccioni, Inail Research. Dept. Risk Assessment, Rome, Italy. Email: a.pelliccioni@inail.it

Table 1: General statistics of observed variables

	CO (mg/m³)		NO (µg/m³)		NO_2 (µg/m³)		O_3 (µg/m³)		T (C°)		RH (%)		GSR (W/m²)	
	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007
Mean	0.61	0.61	20.84	22.32	41.46	43.99	41.55	36.62	13.07	12.97	73.41	73.23	124.15	126.42
SD	0.35	0.38	42.87	41.80	24.70	26.25	41.21	37.46	7.09	7.09	19.23	19.62	219.73	221.95
CV (%)	58.03	61.98	205.70	187.33	59.58	59.67	99.18	102.32	54.25	54.64	26.20	26.80	176.98	175.57
Min	0.1	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	9.0	10.0	0.0	0.0
Max	3.8	4.1	1228.4	398.7	165.3	156.8	227.6	189.1	33.0	37.0	97.0	97.0	1005.0	1002.0
N	8124	8260	7957	8277	7957	8277	8056	8279	8287	8738	8692	8760	8692	8760
Missing	636	500	803	483	803	483	704	481	473	22	68	//	68	//

The ozone concentration follows the maximum of the temperature, especially during the daytime, when the highest values of ozone in urban areas are related to the high values of global solar radiation and pollutants. The analysis of the summer daytime reveals that the ozone peak is at 17.00, the temperature peak is at 15.00 (the temperature values varied in the range of 18° to 26°C) and the GSR peak is at 12.00 for the summer.

The seasonal variation of O_3 shows low concentrations in late autumn and winter and high concentrations in late spring and early summer (average daily 146.94 and 132.3µg/m^3 respectively).

Usually, the distributions of observed pollutants in urban sites are skew, because low values are more frequent than the higher values. This is true for the primary pollutants as well as the secondary ones (Figure 1). Between the two distributions, the main difference consists of the fact that the distribution of primary pollutants presents less skewness than the secondary ones.

This can be explained by the different spatial distribution of the sources.

In general, the asymmetry of the distribution and the identification of outliers (Hawkins, 1980) are very important questions to face for the forecasting models. In our case, ozone distribution is highly skewed (see Figure 1).

In fact, about 97% for 2007 and 96% for 2006 of patterns belong to the class 0-120µg/m^3, whereas less than 0.1% for 2007 and 0.2% for 2006 is above the information threshold (180µg/m^3).

2.2. Methodology

The aim of the study is to suggest a way to implement the learning of neural network models for the ozone forecasting in urban areas.

In general, before running the NN model, the best practice consists of the pre-processing techniques related to the selection of the best information, linked to the pattern selection, to the physics inside your system or to pollutant trends.

The pattern selection is a very important task that should be solved in order to achieve a good generalization of the net, above all if the net is used to simulate chemical reactions in the atmosphere. This task ensures the quality of the data and it optimises the efficiency of the computation time during the elaboration phase and improves the NN learning process.

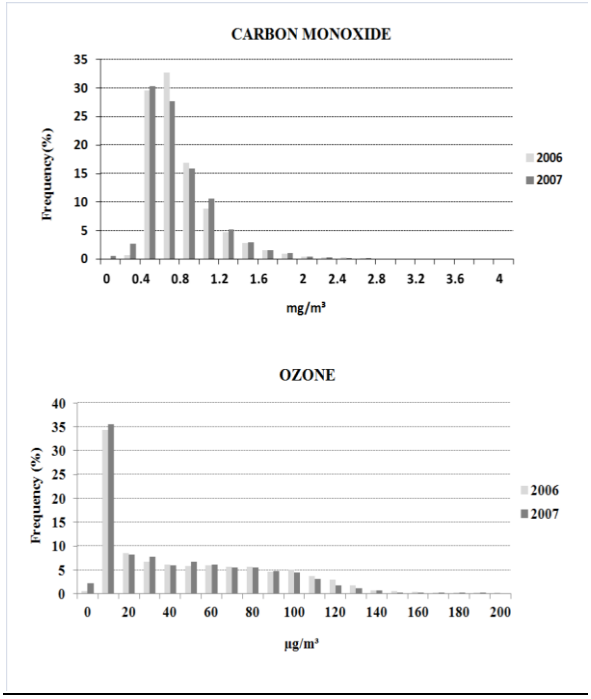

Figure 1: Ozone and CO distribution (2006 and 2007).

We tested the NN performance by means of two ways of pattern selection. The first one is the random pattern selection and the second one is the cluster analysis pattern selection.

2.2.1. Random Pattern selection

The Random pattern selection chooses each unit of dataset that has an equal probability of being in the sample. The selection of each unit is independent of the selection of every other unit. The selection of one unit does not affect the chances of any other pattern. The random pattern selection is the conventional choice for the NN. This method consists in selecting a random sample of patterns in a training set so that the size of the set is reduced by maintaining the same information as the original data set. Its remaining part is used in the generalization phase for evaluating the NN performance. In general, a good training phase can be obtained from 60% to 70% of the total data set. Lower percentage values can be used, but some problems can occur for the test case patterns. Other methods to select patterns, such as the bootstrap (*Franke, 2000)*,(Weigend, 1994), the resampling (Murata, 1994) or cross validation (*Ghiassi, 2005*), *(*Tesauro, 1995) can all be considered as integrated in the random pattern selection scheme.

2.2.2. Cluster Analysis Patterns selection

The novelty of the proposed method to select the environmental patterns consists in the use of the cluster analysis as a discriminate tool for information.

The way to select patterns by using cluster techniques can be seen as an important unsupervised learning technique able to discover the inner structure present in data. In particular, for the ozone, these techniques can be considered very attractive to catch the chemical reactions as well as the pollutant dispersion.

The purpose of cluster analysis is to aggregate a dataset into k separate clusters and to find clusters whose members show a high degree of similarity among themselves, but a high degree of dissimilarity with the members of other clusters. In this way, it is possible to generate a small number of discriminate groups to represent the global information inside the dataset.

Here, the CA is not used to aggregate the information within our data set (i.e. to find the minimum number of cluster reproducing the global variance of all data set). Pattern selection is used as an alternative method to random pattern choice, as a sampling technique, where the entire data set is divided into groups.

The main reason to consider the cluster technique is that is usually more convenient to sample the information on the population in clusters rather than the randomly pattern choice strategy.

The sampling using the cluster methods can be combined with other forms of sampling, for example, random sampling, to ensure that sub-groups are fully represented, or systematic sampling. In our case, we applied the second method because it is more representative and more precise. As systematic sampling we selected only centroids as a representative pattern of each cluster and they are used in the training phase as input of the NN model.

Cluster analysis was conducted by a non-hierarchical method, the k-means technique that can be used to group large numbers of patterns efficiently. The k-means (Kaufman, 1990) is one of the simplest unsupervised algorithms that solve the well known problem of clustering and classifying or grouping a given data set into a k-number of homogeneous groups (clusters).

By utilising the k-means technique, the centroids, as the barycentre of each cluster, are calculated for each pattern. They constitute the new dataset, composed with the centroid of each cluster, which is divided into a training set used for constructing the predictor and one test set for evaluating its performance. The subsets of the new associated training and generalisation patterns are built up and used to test and validate patterns

2.2.3. Neural Network

The greatest advantage of a neural network is its ability to model a complex non-linear relationship between input and output variables (Gardner, 1999), (Gardner, 2000), (Abdul-Wahab, 2002), such as those in environmental systems, without a priori assumptions about its nature (the implicit properties of the model, its parameters and the observed variables) and data distribution from which the modelling sample is drawn. The selection of an appropriate network topology depends on the number of parameters, the weights, the selection of an appropriate training algorithm and the type of transfer functions used.

For the transfer function, we use the most common architecture, the Multi Layer Perceptron (MLP) (Fausett, (1994), Ripley, (1996), Rojas (1996)), which is a type of feed-forward of the neural network and generally uses the back-propagation algorithm to develop a model to illustrate relationships between inputs and the desired output for the training data.

The inputs are fully connected to the first hidden layer. Each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs.

We tested different NN parameters in order to choose the best model. After several simulations, running for 3000 epochs and testing different numbers of hidden layers (9, 6 and 12), we selected 12 hidden layers as the best performance of the perceptron network.

In this context, we used the MLP model with a single hidden layer, with 12 hidden neurons and with the following choice for the sigmoid activation function for the hidden units:

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

It can be observed that the activation function (Equation 1) is able to optimise the learning process in the best way with respect to the conventional choice, where A=1 and K=1.

The input layer contains the main and essential variables for the ozone, such as the hourly CO, NO, NO_2, T, RH and SR, whereas the target neuron is the ozone itself. Our final NN is constituted by 6-12-1 neurons, where the hidden neurons come from the above elaborations.

3. Results

We applied NN to the results coming from the pattern selection process to forecast ozone concentrations using as input data, meteorology,

as well as primary and secondary pollutants (CO, NO, NO2). We carried out 27 simulations using different percentages of input patterns for the training. All the results are referred to the generalization phase, where the patterns are never seen by the NN .

The results obtained by Cluster Analysis applied to the NN (named CANN) are compared to the Conventional Random Pattern Selection applied to the NN (named CRPSNN), our benchmark, with different percentages of input patterns from 0.01% to 100%.

In general, we observed that the NN performance shows different values for the ozone predictions. In particular, Figure 2 shows that CANN is performed and predicted better than CRPSNN. In terms of global fit, CANN has a better performance (R^2 from 0.59 to 0.89) than CRPSNN (R^2 from 0.05 to 0.97). These results show a meaningful difference and demonstrate that the pre-selection by cluster can simulate in best way the physics of ozone (i.e. can reproduce the observed skew distribution for ozone). Moreover, we observed that using only 1% of total data we obtain R^2=0.56 utilizing CANN, whereas we obtain R^2=0.41 for CRPSNN (Figure 2). The first important result of our elaboration shows that CA sampling is, on average, more efficient than CRPSNN when using small amounts of patterns during the training and, consequently, could be adapted to simulate rare events, such as what happens during the extreme ozone episodes.

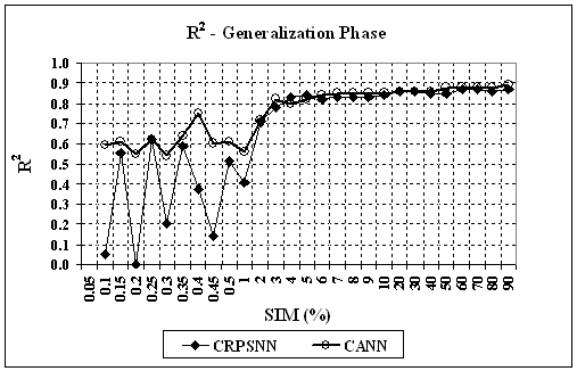

Figure 2: R^2 performance

If we consider the rate between R^2 calculated by generalisation and by the training phase (Rate of Determination-RD), we observe another interesting result. Usually, the conventional way to train the NN presents a marked maximum value of this RD corresponding to the well-known decreasing of the generalisation performance of NN with the increase of percentage of input training data (the so called over-fitting question).

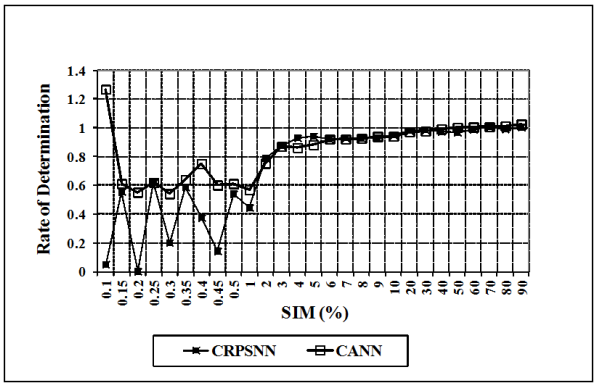

Figure 3: Trend of Rate of determination at different simulations

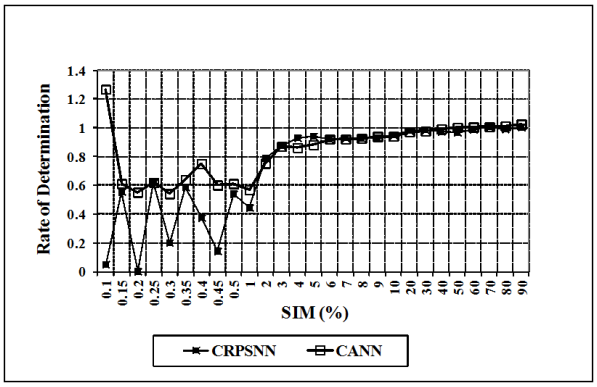

Figure 3 shows the RD calculated by our simulations. The maximum and stable performance for the generalisation is verified with 30% and 40% of input data for both approaches.
While in the random selection by CRPSNN the performance decreases in a meaningful way when we move from this percentage of input patterns (note the value of 0.51 if we use 0.2% of data), in the selection by cluster sampling the performance can be considered satisfactory in the generalisation cases. The values of RD greater than 1 for 0.10% demonstrate that R^2 in generalization is a little greater than the training ones. These results could be linked to the coincidences of the number of patterns with the optimisation of variability (variance within and between) of each cluster.

4. Conclusions

Our research shows a good capacity of the NN to analyse the large complex data sets and to model the ozone levels using the clustering approach during pattern pre-processing phase.
The capability of the Neural Network technique, applied to multivariate and non-linear problems, to capture the environmental information inside the data depended not only on the learning methods used, but also on the preliminary study of patterns. This paper is related to the quality of the data used to train the neural network. The

problem of pre-processing and of a good sampling plan for the input data is essential to obtain a good forecasting performance of the NN. We observed that the neural classifier trained after the random pattern choice, is able to distinguish only average/stable situations. On the contrary, the NN, after cluster pattern choice, is able to distinguish outlier situations too.

In conclusion, the clustering technique, adopted as a pattern selection approach, obtains better predictions of pollutant phenomena. The results are very encouraging and our simulations based on cluster analysis demonstrated that this method is feasible and effective, resulting in a substantial reduction of data input requirement and outperform other techniques applied in this context. The determination coefficient (RD) substantially shows better performance in the combined forecast procedure.

REFERENCES

Abdul-Wahab S.A., Al-Alawi, S.M., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software 17, 219-228.

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

EPA, 2006. Air Pollutants. Epa.gov. 2006-06-28. http://www.epa.gov/ebtpages/airairpollutants. html. Retrieved 2010-08-29.

Fausett L., 1994. Fundamentals of Neural Networks. In: Architectures, Algorithms and Applications. Prentice Hall, Englewood Cliffs, NJ 07632.

Finlayson-Pitt JB, Pitts WJ, 1986. Fundamental and Experimental Techniques. Atmospheric chemistry. John Wiley and Sons, Inc.,New York, Brisbane, Toronto, Singapore, pp 108, 136

Franke, M. H. Neumann,2000. Bootstrapping Neural Networks. Journal Neural Computation Volume 12 Issue 8, August 2000,pp 1929 – 1949, MIT Press Cambridge, MA, USA

Gardner M.W, Dorling S.R., 1998. Artificial Neural Networks (the Multilayer Perceptron)-E Review of applications in the atmospheric sciences, Atmos. Environ., 32(14/15), 2627-2636.

Gardner M.W., Dorling, S.R., 1999: Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment 33, 709-719.

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

Ghiassi M.,Saidane, H., Zimbra, D.K.,2005. A dynamic artificial neural network model for forecasting time series events. International Journal of Forecasting, 21(2) pp 341–362

Kaufman, L. and P. J. Rousseeuw, 1990. Finding Groups in Data. An Introduction to Cluster Analysis. John Wiley and Sons.

Hawkins D., 1980. "Identification of Outliers". Chapman and Hall, London.

Murata N., Yoshizawa S., Amari S.,1994. Network information criterion-determining the number of hidden units for an artificial neural network model. Neural Networks, IEEE Transactions on, , 5(6) pp 865 – 872

Pelliccioni A., Lucidi S., La Torre V.,Pungì F.,2010 a, "Optimization of Neural Network performances by means of exogenous input variables for the forecast of Ozone pollutant in Rome Urban Area" Eighth Conference on Artificial Intelligence and its Applications to the Environmental Sciences" , AMS 90th Annual Meeting 17–21 January 2010.

Pelliccioni A., Cotroneo R., Pungi F., 2010 b, "Optimization of neural net training using patterns selected by cluster analysis: a case-study of ozone prediction level" Eighth conference on Artificial Intelligence and its Applications to the Environmental Sciences, AMS 90th Annual Meeting, Atlanta, Georgia

Ripley B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.

R. Rojas, 1996. Neural Networks, Springer-Verlag, Berlin.

Seinfeld, J.H., 1986. Atmospheric chemistry and physics of air pollution. John Wiley and Sons, Inc.,Somerset, NJ.

Tesauro G., Touretzky D.S., Leen T.K.,1995. Advances in Neural Information Processing Systems. Edizione 7, MIT Press, 1

Weigend S.,Le Baron B.,.1994 Evaluating Neural Network Predictors by Bootstrapping. ICONIP : International Conference On Neural Information Processing, pp. 1207-1212.

World Health Organisation, 2003. Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide