# OPTIMIZATION OF NEURAL NET TRAINING USING PATTERNS SELECTED BY CLUSTER ANALYSIS: 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.

## INTRODUCTION

In the atmosphere ozone is a well known secondary air pollutant, results from complex chemical reactions and is a very reactive gas and presents concentration levels which are strongly dependent both from the micro-meteorological conditions and the seasonal effects. The prediction of ozone ($O_3$) levels is very hard to find as evident in different works [1] [2] [3] [4] [5]. The present work aims at predicting the Ozone levels [6] in the urban area of Rome using neural networks (NN) as model and utilizing a novel strategy for choosing of input patterns.
During the training phase [7], we used cluster analysis techniques (K-means algorithm), in order to optimise the selection of input patterns.
In NN training phase, usually the main problems concern the represantive pattern selection to be used in the generalization perform, as well as variables distribution representative of all information.
As known, the performance of generalisation is highly dependent by the significance of pattern selection. In general, during the training the selection involve a random pattern choice starting from some percentage of total data. In our work, we apply cluster analysis (CA) for the patterns selection during the training phase. This approach improves the accuracy of the ozone prediction, enhances the learning capabilities and NN potential to predict ozone and, a very interesting result, synthesizes in correct way information for large data set.

## DATA SET DESCRIPTION

Our time series is derived from a background monitoring station of the ARPAL (Environmental Protection Agency of Lazio Region) in Rome (Villa Ada monitoring station), during all the calendar year 2007. We have considered ozone one-hour concentrations from monitoring stations positioned in Rome which have had for each year at least 50% valid data (taking count of EoI – Exchange of Information). The city of Rome is characterized by frequent ozone peaks, associated with hot sunny days and turbulence conditions. Other important factors derive by the main primary pollutants (NO, $NO_2$, CO) coming from the main urban sources. Villa Ada monitoring station represents typical sub-urban situations with high ozone concentration levels located in the NNW direction.

Our dataset concerns about 7370 hourly patterns and is composed by pollutants variables and conventional meteorological variables. It was decided that this study would use a conservative number (four) of pollutant (concentrations of ozone and other relevant pollutants) and meteorological variables in order to maintain parsimony and keep the resulting models simple enough for meaningful comparison. In Table 1, we are described these variables.

**Table 1: Variables Dataset**

| Pollutants | | | Meteorological | | |
|---|---|---|---|---|---|
| Carbon monoxide | $mg/m^3$ | CO | Temperature | C° | T |
| Nitrogen Oxide | $\mu g/m^3$ | NO | Relative Humidity | % | RH |
| Nitrogen Dioxide | $\mu g/m^3$ | $NO_2$ | Pressure | mbar | P |
| **Ozone (Output)** | $\mu g/m^3$ | **O3** | Rain | mm | |
| | | | Global Solar Radiation | $W\backslash m^2$ | GSR |

Global radiation were analysed to investigate the ozone correlation with photochemical reactions .
Moreover, our time series show realistic ozone patterns concerned about 8760 hourly average data for each city. All data were previously standardised, before conducting any analysis.

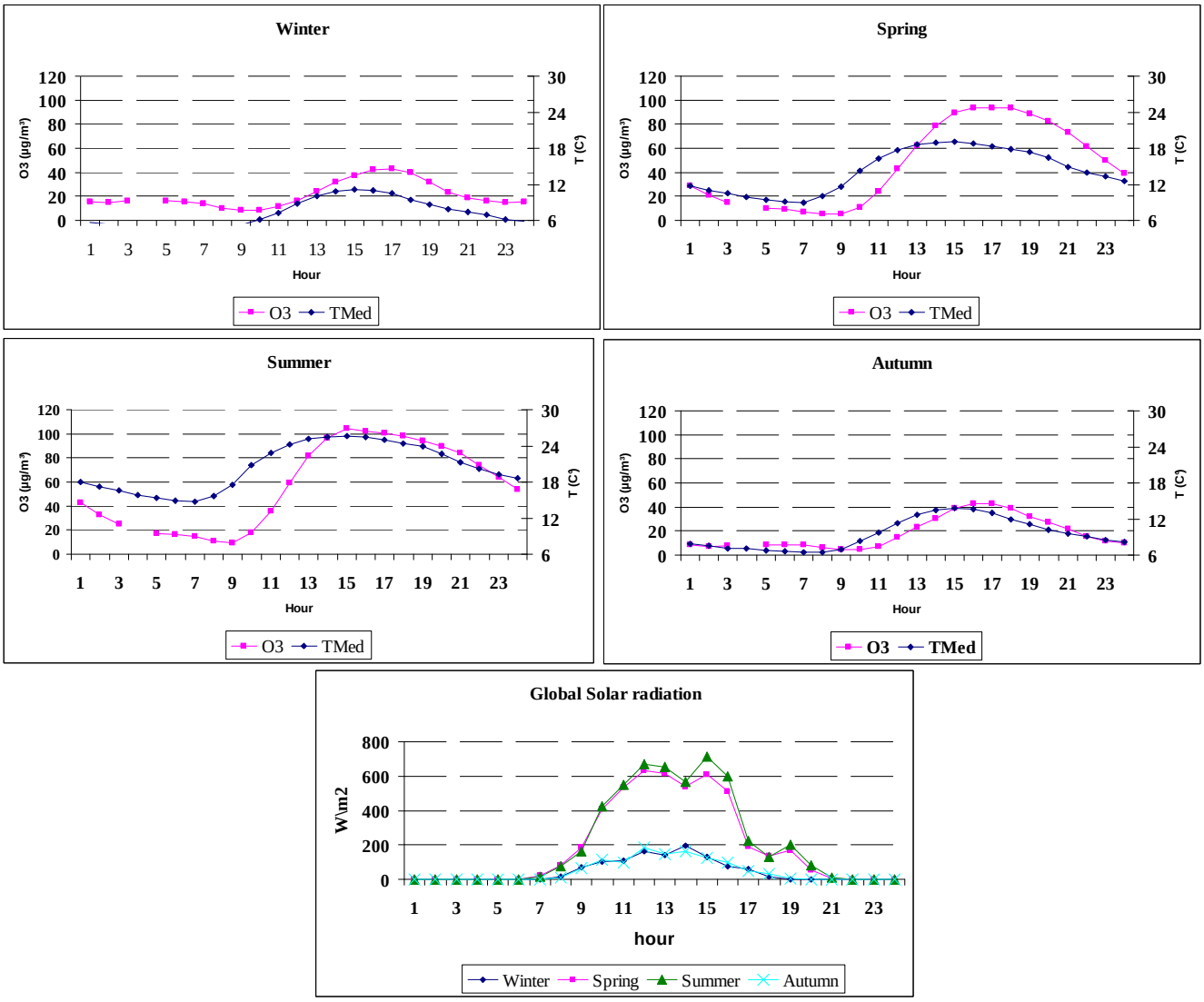
Table 2 shows the general statistics calculated for the pollutants and meteorological parameters used in this study. We analyze time series of the following statistical characteristics of the distribution of hourly data: mean, standard deviation, maximum and minimum ozone values for all seasons. We observed that the maximum hourly ozone per year lies around a value of 189.1 µg/m³ and a considerable variability in the time series. In our environmental time series analysed, we have missing values that depend on different working periods of the stations or occasional malfunctions of the instruments. This analysis cannot account for any missing values.

**Table 2: General statistics**

| | CO (µg/m³) | NO (µg/m³) | NO2 (µg/m³) | O3 (µg/m³) | T (C°) | RH (%) | Press (mbar) | WS (m/s) | Rain (mm) | GSR ($W/m^2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.61 | 22.32 | 43.99 | 36.62 | 12.97 | 73.23 | 1016.34 | 0.78 | 0.07 | 126.42 |
| Standard deviation | 0.38 | 41.80 | 26.25 | 37.46 | 7.09 | 19.62 | 6.25 | 0.79 | 0.49 | 221.95 |
| CV (%) | 62.30 | 187.28 | 59.67 | 102.29 | 54.66 | 26.79 | 0.61 | 101.28 | 700.00 | 175.57 |
| Min | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 | 10.0 | 987.0 | 0.0 | 0.0 | 0.0 |
| Max | 4.1 | 398.7 | 156.8 | 189.1 | 37.0 | 97.0 | 1038.0 | 5.9 | 12.6 | 1002.0 |
| N | 8260 | 8277 | 8277 | 8279 | 8738 | 8760 | 8760 | 8755 | 8760 | 8760 |
| Missing | 500 | 483 | 483 | 481 | 22 | | | 5 | | |

Figure1 shows a relationship during the time of a typical day between temperature, global solar radiation and ozone obtained considering the mean values along each season. Daytime usually is linked to the main turbulence conditions related to solar elevation, geographical positions, seasonal effects. We note that the ozone shows high concentrations (104.4 µg/m³)during daytime and low concentrations (7.3µg/m³). during late night and early morning (4.3 µg/m³).
We observed that the influence of temperature on the ozone concentration values was examined based on temporal fluctuations. The ozone concentration distribution follows the maximum temperature , especially during the daytime, when the highest values of ozone in urban area are related to the high values of solar radiation and pollutants. We observed that the ozone peak is at 15 p.m., temperature peak is at 15 p.m. (The temperature values varied in the range of 18 to 26C°) and GSR peak at and 12 p.m. for summer season.

**Figure 1: Mean hourly temperature, global solar radiation and ozone (2007)**




We observed (Table 3, Table 4 and Table 5) that the T, GSR and O3 are distributed according to each month of the year from all hours of day. An idea of the diurnal variation of the GSR, T and O3 in the different seasons was given by the seasonal variation of the ratio daily/month of these variables, as well as by information on the highest and lowest values of GSR, T and O3, depending on the hourly measurements. In particular, we observed high level of ozone (119.6 µg/m³), temperature (27.2 C°) and global solar radiation (863.6 W/m$^2$) during the daytime and spring-summer season and low level during nighttimes and autumn-winter season.

**Table 3: Mean hourly ozone for month (µg/m³)**

| Hour\ Months | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 13.2 | 12.8 | 23.0 | 19.3 | 36.1 | 46.5 | 44.9 | 44.4 | 27.4 | 9.7 | 5.8 | 5.2 |
| 2 | 13.6 | 11.4 | 25.6 | 10.3 | 24.1 | 38.4 | 28.8 | 38.1 | 20.1 | 8.7 | 6.3 | 4.2 |
| 3 | 14.8 | 12.7 | 27.7 | 5.9 | 13.7 | 28.6 | 17.1 | 31.5 | 19.2 | 10.1 | 6.4 | 4.4 |
| 4 | | | | | | | | | | | | |
| 5 | 14.4 | 11.6 | 28.6 | 3.9 | 7.4 | 18.5 | 9.5 | 23.5 | 14.1 | 10.3 | 6.4 | 6.2 |
| 6 | 14.1 | 11.4 | 26.5 | 2.6 | 8.0 | 15.1 | 9.3 | 23.1 | 14.1 | 10.0 | 6.3 | 6.3 |
| 7 | 11.7 | 10.6 | 23.4 | 2.4 | 7.8 | 11.9 | 9.0 | 21.8 | 11.5 | 10.2 | 5.6 | 6.6 |
| 8 | 9.1 | 6.9 | 17.0 | 1.0 | 6.8 | 10.4 | 6.0 | 17.5 | 7.1 | 5.9 | 4.5 | 6.2 |
| 9 | 8.9 | 5.1 | 13.2 | 0.5 | 7.4 | 11.9 | 6.7 | 13.6 | 5.1 | 4.1 | 3.1 | 4.4 |
| 10 | 9.3 | 3.5 | 16.0 | 3.3 | 14.2 | 20.1 | 17.6 | 22.1 | 8.0 | 4.2 | 3.4 | 3.5 |
| 11 | 10.3 | 4.0 | 25.1 | 13.2 | 30.7 | 37.3 | 38.9 | 40.7 | 20.3 | 7.4 | 5.6 | 4.0 |
| 12 | 12.3 | 9.2 | 37.7 | 29.8 | 50.9 | 56.6 | 70.0 | 60.3 | 42.0 | 15.0 | 13.1 | 6.7 |
| 13 | 16.9 | 18.1 | 50.8 | 54.8 | 69.0 | 74.8 | 95.5 | 81.9 | 61.3 | 24.7 | 20.1 | 11.8 |
| 14 | 21.3 | 26.2 | 63.2 | 75.3 | 82.8 | 89.5 | 113.0 | 95.7 | 74.9 | 35.4 | 24.2 | 15.6 |
| 15 | 25.7 | 31.9 | 69.8 | 91.5 | 91.8 | 98.9 | 119.6 | 103.1 | 86.0 | 48.9 | 30.2 | 18.5 |
| 16 | 31.1 | 36.8 | 74.3 | 97.7 | 95.4 | 100.7 | 111.0 | 101.2 | 90.1 | 55.4 | 33.0 | 21.3 |
| 17 | 30.2 | 40.1 | 75.0 | 100.5 | 95.7 | 96.6 | 106.5 | 100.0 | 92.5 | 58.6 | 30.9 | 17.7 |
| 18 | 25.3 | 39.4 | 72.0 | 100.7 | 96.2 | 92.5 | 104.0 | 99.2 | 90.0 | 56.5 | 25.0 | 11.9 |
| 19 | 17.2 | 31.5 | 65.5 | 94.0 | 92.8 | 86.3 | 102.1 | 96.2 | 84.5 | 48.9 | 15.8 | 8.5 |
| 20 | 12.5 | 20.3 | 55.3 | 85.7 | 88.5 | 79.8 | 98.3 | 92.0 | 79.4 | 45.2 | 9.7 | 6.9 |
| 21 | 12.0 | 16.1 | 42.4 | 73.4 | 82.9 | 75.5 | 92.1 | 87.0 | 70.9 | 35.8 | 6.2 | 4.5 |
| 22 | 12.1 | 13.3 | 33.8 | 57.2 | 72.2 | 68.9 | 81.3 | 78.5 | 56.2 | 25.2 | 5.1 | 4.0 |
| 23 | 11.1 | 12.2 | 30.7 | 42.8 | 59.0 | 60.3 | 67.3 | 68.2 | 47.9 | 16.9 | 4.0 | 4.7 |
| 24 | 11.9 | 14.4 | 26.7 | 28.8 | 49.0 | 54.5 | 56.8 | 55.7 | 37.3 | 12.2 | 5.0 | 5.4 |

**Table 4: Mean hourly temperature for month (C°)**

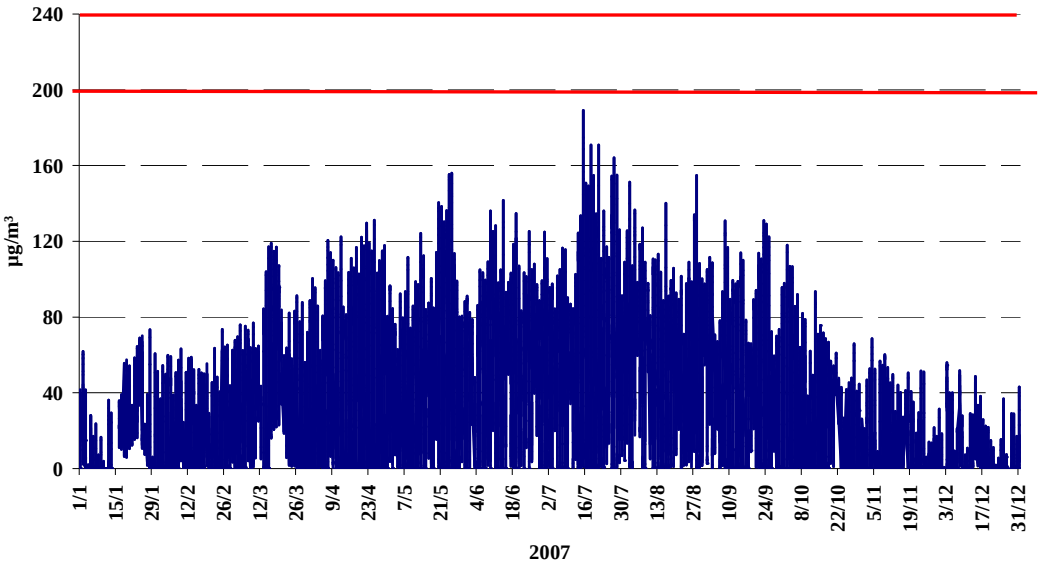
| Hour \ Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.2 | 5.4 | 6.8 | 10.2 | 13.0 | 16.8 | 18.4 | 19.2 | 14.7 | 10.9 | 6.0 | 4.0 |
| 2 | 4.9 | 5.0 | 6.5 | 9.4 | 12.3 | 16.2 | 17.7 | 18.5 | 14.1 | 10.5 | 5.6 | 3.8 |
| 3 | 4.8 | 5.1 | 6.1 | 8.8 | 11.6 | 15.7 | 16.7 | 17.8 | 13.6 | 10.2 | 5.1 | 3.5 |
| 4 | 4.8 | 4.9 | 5.9 | 8.3 | 11.0 | 15.1 | 15.8 | 17.2 | 13.1 | 9.5 | 5.2 | 3.8 |
| 5 | 4.5 | 4.6 | 5.5 | 7.7 | 10.5 | 14.4 | 15.2 | 16.6 | 12.6 | 9.4 | 4.9 | 3.5 |
| 6 | 4.7 | 4.5 | 5.3 | 7.3 | 10.3 | 14.1 | 14.9 | 16.3 | 12.2 | 9.2 | 4.9 | 3.6 |
| 7 | 4.6 | 4.1 | 5.2 | 7.0 | 10.2 | 14.3 | 14.8 | 15.8 | 11.9 | 8.8 | 4.7 | 3.5 |
| 8 | 4.7 | 4.2 | 5.7 | 7.5 | 11.6 | 15.6 | 16.3 | 16.8 | 12.1 | 8.7 | 4.8 | 3.3 |
| 9 | 4.6 | 4.9 | 7.0 | 9.3 | 13.4 | 17.3 | 18.5 | 18.5 | 13.6 | 9.4 | 5.5 | 3.1 |
| 10 | 4.9 | 5.8 | 8.9 | 12.5 | 15.8 | 19.9 | 22.2 | 21.5 | 16.5 | 11.3 | 6.5 | 3.6 |
| 11 | 6.0 | 7.3 | 10.1 | 14.9 | 17.6 | 21.8 | 24.5 | 23.3 | 18.6 | 13.2 | 8.1 | 4.1 |
| 12 | 7.5 | 8.8 | 11.6 | 16.7 | 18.7 | 23.2 | 26.3 | 24.8 | 19.7 | 14.8 | 9.9 | 5.5 |
| 13 | 8.9 | 9.9 | 12.4 | 17.7 | 19.5 | 23.9 | 27.2 | 25.9 | 20.8 | 16.6 | 10.8 | 7.1 |
| 14 | 9.8 | 10.6 | 13.0 | 18.2 | 19.6 | 24.2 | 27.2 | 26.3 | 21.2 | 17.4 | 11.6 | 8.0 |
| 15 | 10.2 | 10.9 | 13.2 | 18.3 | 20.0 | 24.2 | 26.9 | 26.7 | 22.0 | 18.0 | 11.6 | 8.3 |
| 16 | 10.2 | 10.8 | 12.6 | 18.1 | 19.7 | 23.8 | 26.8 | 26.4 | 21.8 | 17.9 | 11.6 | 8.2 |
| 17 | 9.7 | 10.6 | 12.2 | 17.2 | 19.4 | 23.5 | 26.3 | 25.7 | 21.0 | 17.6 | 10.6 | 7.5 |
| 18 | 8.5 | 9.6 | 11.2 | 16.7 | 19.1 | 22.9 | 25.8 | 25.2 | 20.3 | 16.7 | 9.5 | 6.3 |
| 19 | 7.8 | 8.8 | 10.3 | 15.8 | 18.8 | 22.7 | 25.5 | 24.6 | 19.5 | 15.3 | 8.8 | 5.8 |
| 20 | 7.1 | 8.1 | 9.5 | 14.7 | 17.9 | 21.6 | 24.5 | 23.4 | 18.2 | 14.2 | 7.8 | 5.2 |
| 21 | 6.6 | 7.6 | 8.8 | 13.4 | 16.3 | 20.0 | 22.7 | 22.0 | 17.3 | 13.5 | 7.3 | 4.4 |
| 22 | 6.2 | 7.1 | 8.4 | 12.5 | 15.2 | 18.8 | 21.2 | 21.1 | 16.7 | 12.9 | 6.6 | 4.2 |
| 23 | 5.6 | 6.4 | 7.8 | 11.8 | 14.5 | 18.0 | 20.3 | 20.3 | 16.0 | 12.3 | 6.1 | 3.6 |
| 24 | 5.5 | 6.0 | 7.1 | 11.1 | 13.7 | 17.5 | 19.5 | 19.6 | 15.1 | 11.5 | 6.0 | 3.8 |

**Table 5: Mean hourly global solar radiation for month (W/m$^2$)**

| Hour \ Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6 | 0.9 | 0.4 | 0.3 | 0.2 | 0.2 | 0.0 | 0.1 | 0.3 | 0.5 | 0.6 | 0.8 |
| 2 | 0.6 | 0.8 | 0.5 | 0.3 | 0.3 | 0.3 | 0.0 | 0.2 | 0.4 | 0.6 | 0.7 | 0.7 |
| 3 | 0.9 | 1.0 | 0.7 | 0.3 | 0.4 | 0.2 | 0.0 | 0.2 | 0.2 | 0.6 | 0.9 | 0.8 |
| 4 | 1.1 | 1.0 | 0.6 | 0.2 | 0.4 | 0.2 | 0.1 | 0.2 | 0.2 | 0.7 | 1.0 | 0.6 |
| 5 | 1.3 | 1.0 | 0.7 | 0.3 | 0.5 | 0.4 | 0.2 | 0.2 | 0.1 | 0.8 | 1.0 | 0.7 |
| 6 | 1.0 | 1.0 | 0.7 | 0.3 | 2.2 | 3.9 | 1.5 | 0.2 | 0.3 | 0.6 | 0.9 | 1.0 |
| 7 | 1.0 | 1.0 | 5.8 | 8.1 | 28.2 | 37.1 | 28.1 | 10.6 | 1.6 | 0.7 | 1.4 | 1.0 |
| 8 | 4.4 | 13.3 | 52.5 | 63.4 | 91.0 | 100.1 | 89.8 | 78.3 | 41.2 | 10.1 | 16.0 | 4.9 |
| 9 | 41.4 | 63.4 | 154.2 | 154.8 | 199.2 | 219.4 | 189.7 | 135.9 | 140.2 | 59.2 | 70.3 | 43.8 |
| 10 | 75.4 | 81.3 | 204.4 | 376.3 | 467.3 | 473.0 | 538.9 | 389.4 | 275.4 | 141.9 | 74.8 | 77.9 |
| 11 | 84.3 | 101.9 | 210.1 | 508.9 | 573.0 | 634.6 | 686.6 | 533.6 | 303.1 | 109.1 | 89.3 | 69.7 |
| 12 | 106.2 | 173.3 | 306.5 | 649.1 | 662.7 | 722.5 | 787.1 | 647.9 | 438.6 | 226.0 | 161.5 | 113.0 |
| 13 | 137.2 | 128.1 | 203.8 | 549.2 | 687.8 | 761.6 | 837.3 | 685.4 | 272.7 | 161.1 | 128.6 | 143.3 |
| 14 | 176.5 | 167.1 | 250.8 | 350.7 | 679.5 | 778.2 | 863.6 | 496.2 | 148.6 | 124.5 | 183.2 | 198.5 |
| 15 | 91.7 | 123.5 | 262.6 | 584.5 | 662.9 | 742.9 | 834.2 | 681.9 | 476.4 | 183.1 | 67.4 | 63.6 |
| 16 | 67.0 | 79.8 | 146.0 | 483.7 | 558.5 | 669.4 | 744.7 | 591.5 | 325.2 | 114.3 | 70.5 | 59.1 |
| 17 | 30.5 | 71.5 | 105.1 | 139.6 | 189.9 | 332.6 | 323.3 | 180.0 | 93.2 | 89.3 | 22.3 | 12.4 |
| 18 | 1.6 | 15.3 | 57.8 | 133.7 | 127.9 | 171.4 | 124.8 | 121.5 | 135.9 | 72.9 | 0.6 | 0.5 |
| 19 | 0.6 | 0.5 | 15.6 | 122.1 | 204.3 | 258.6 | 285.5 | 180.8 | 75.2 | 9.3 | 0.5 | 0.7 |
| 20 | 0.5 | 0.5 | 1.9 | 24.4 | 75.9 | 112.8 | 143.7 | 57.6 | 7.2 | 0.5 | 0.5 | 0.6 |
| 21 | 0.8 | 0.5 | 0.5 | 0.4 | 5.0 | 18.9 | 16.1 | 2.1 | 0.1 | 0.5 | 0.5 | 0.7 |
| 22 | 0.8 | 0.6 | 0.5 | 0.2 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.5 | 0.5 | 0.6 |
| 23 | 0.6 | 0.6 | 0.4 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.4 | 0.5 | 0.6 |
| 24 | 0.6 | 0.5 | 0.2 | 0.3 | 0.2 | 0.1 | 0.0 | 0.1 | 0.1 | 0.4 | 0.6 | 0.8 |

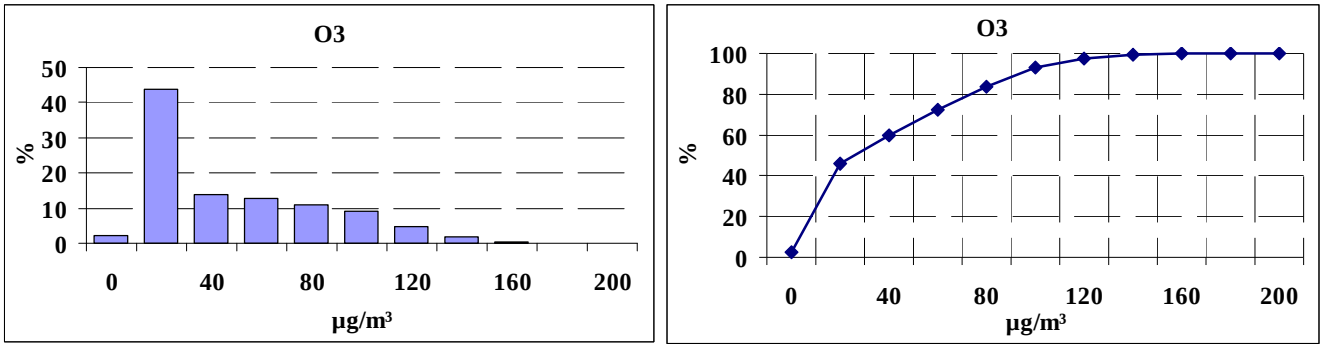
The Figure 2 shows ozone trend for the 2007. The seasonal variation in O3 shows low concentrations in late autumn and winter and high concentrations in late spring and early summer (189.1 µg/m$^3$). In Figure 2 are also given typical alarm ozone levels.

**Figure 2: Time series plot of Ozone**




Usually, the pollutants distribution presents a skewness and the identification of outlier [8] situations are one of the more important problems. In our case, Ozone's distribution is highly skew (Figure 3). In fact, about 97% of patterns belonging to the class 0-120 µg/m$^3$, while less than 0.1% is above the information threshold (180 µg/m$^3$).

**Figure 3: Ozone distribution**




## METHODOLOGY

Different methodologies could be employed to optimise NN performances. As known [ 9 ], one of the main weakness factors regards the meaningful of the pattern choice during the NN training phase as regards generalisation one, and consequently we concentrate our attention on patterns selection techniques in order to selection the best patterns.
Patterns selection used for artificial neural network training phase is one of the most important tasks that should be solved in order to achieve good generalisation of the net.
In general, the pattern choice was executed by the random pattern selection for different percentage of input data. In NN training, all patterns of the main dataset are usually presented equally probability of being selected for the partition dataset, often in random order. In other words, this method consists of selecting a random subset of patterns selection into a training set so that the size of the set is reduced while its representativeness power is not affected. The complement data are used in the generalization phase for evaluating its performance.
In generally, people use CA technique to reduce the number of meaningful data without any loss of information. This technique was adopted in different contest to select the best during NN training phase.
In our approach, training pattern selection procedure is given by the cluster techniques (K-means algorithm), that is an important technique used in discovering some inherent structures present in data and does not require further assumptions or a priori knowledge (pre-clustered). The purpose for the partitioning of a dataset of objects into k separate clusters is 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 groups to represent (summarize) the dataset.
CA techniques could solve pattern classification problems related to NN, simplifying and selecting the best patterns from dataset, and so could improve intelligence to the NN models. So doing, we intend to suggest a method for the choice of patterns that is able to optimize the NN training with a small amount of input patterns, to simulate in urban area, like Rome, chemical reactions for the Ozone levels and to simulate outlier situations(i.e. high hourly ozone peaks).
In our work, cluster analysis is used in not conventional way and it was not interested to the significant information typical of clustering, but in the percentage of well selected patterns to use during the training.
We utilise cluster methods exclusively to select the best and significant patterns, while these techniques usually are used to synthesise data in homogeneous group whose average dissimilarity to all other items in the same cluster is minimal.
CA was conducted by not hierarchical method, k-means technique that can be used to group large number of patterns efficiently.
The K-means [10] is one of the simplest unsupervised algorithms that solve the well known clustering problem and classify or group a given data set into K number of homogeneous groups (clusters) with respect to the compositional behaviour in the temporal domain The grouping is done by minimizing the sum of squares of distances between data and the corresponding centroids, which defines the geometric centre of the cluster. (see equation 1).

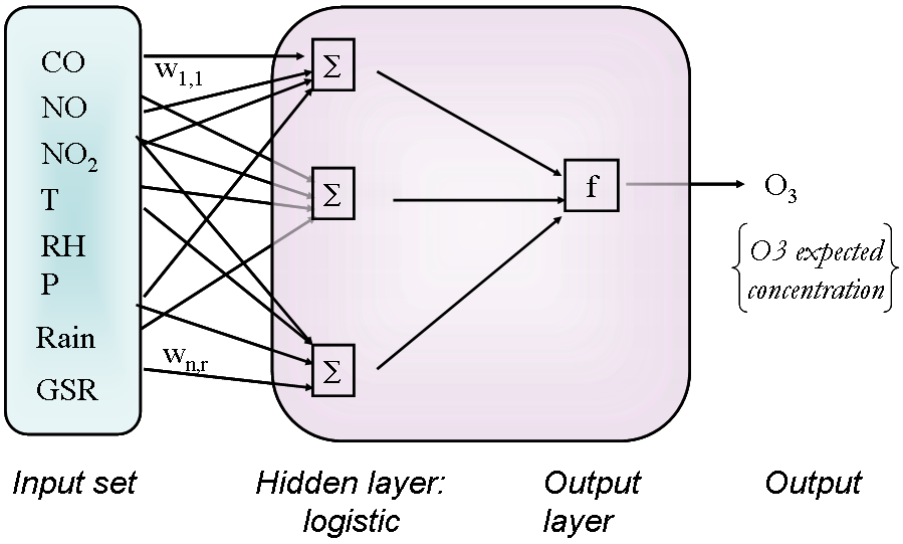
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre $c_j$, is an indicator of the distance of the n data points from their respective cluster centres.
In CA application, the aim of our study regards the choice of the best input pattern for NN. In this contest, we don't search the conventional optimal number of groups but the best selection of patterns for NN more representative than random patterns choice. So, we decided to compare this two pattern choice during NN training. We fix same percentages of input patterns from 1% up 90%. We select these percentages using random pattern choice and CA technique.
Centroids are calculated for each pattern and constitute the new dataset to apply to the NN, divided into a training set used for constructing the predictor and one test set for evaluating its performance.The new training and generalisation patterns subsets associated are built up are used as test and validation patterns, respectively. Pattern centroids constitute the new significant dataset used during the training phase (the training-set centroids). In this phase, the training-set centroids are used to reduce the amount of patterns to be learned for the neural network and to optimize the NN training phase. In fact, the centroids can minimize the mean-squared error of our original dataset.
We train the NN using the patterns constituted by the centroids coming from k-means algorithm; using ad hoc percentages of the whole dataset (we tested 1% up to 90% of total data amount).
After the neural network has been successfully trained, its performances are tested on a separate testing sets constituted by the original, that did not contain centroids. In this way, it is possible to verify higher accuracy of generalisation and prediction of our approach than one trained with patterns drawn from centroids dataset.
We observe that combination of Neural Net architecture in conjunction with a robust number of centroids improves system performance. This is the real novelty of our approach.
The results of our approach are compared with the Conventional Random Pattern Selection (CRPS), our benchmark, for different percentage of input patterns.

As NN architectures, we use the Multi Layer Perceptron (MLP) [11] [12] [13] model with a single hidden layer, 10 hidden neurons and with sigmoid activation function (see equation 2) that approximates nonlinearities.

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (2)$$

**Figure 4: MLP architecture**



We tested different numbers of neurons in the hidden layer (12 and 14 hidden neurons), but the best performance of perceptron network was obtained by 10 neurons. The choice of 10 hidden neurons is based on two considerations: maximizing the hidden neurons to increment the NN parameters and simultaneously minimizing this number in relation with the main situations linked to the input patterns. Moreover, we utilise different neurons methods to optimize the weight values, in the manner that the errors of the network output ca by searching optimal directions n be minimized, and our results derived by conjugate gradient algorithm, that can be used to minimize network output error and accelerated the convergence rate by searching optimal solutions. For this algorithm, the computation time is proportional to the number of weights selected. As note, conjugate gradient does not require the user to specify learning rate and momentum parameters.

**Table 6. Neural Networks architecture**

| | |
|---|---|
| NEURAL NETWORK MODEL | MLP 12-10-1 |
| HIDDEN NEURONS | 10 |
| ALGORITHM | CONJUGATE GRADIENT |
| EPOCH | 3000 |
| ERROR FUNCTION | SUM OF SQUARE |
| HIDDEN ACTIVATION FUNCTION | LOGISTIC |
| OUTPUT ACTIVATION FUNCTION | IDENTITY |
| NETWORK RANDOMIZED | NORMAL |

## RESULTS AND DISCUSSION

We applied NN to the results coming from the patterns selection process to forecast time series of ozone levels concentrations using as input data, meteorology, as well as primary and secondary pollutants (CO, NO, $NO_2$).
The performance of our approach is compared with CRPS in term of determination coefficient ($R^2$) and the rate of efficiency for different percentage of input patterns from 1% to 100% (from Test 1 to Test 14), excluding negative Ozone concentrations predicted by NN during the generalizations phase.

**Table 7: NN RESULT**

| | Training Patterns | | Test Patterns | | CLUSTER NN | | | CRPS NN | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $R^2$ | | Neg O3 prediction | $R^2$ | | Neg O3 prediction |
| | N | % | N | % | Train | Gen | Gen (%) | Train | Gen | Gen (%) |
| **Test 1** | 74 | 1 | 7296 | 99 | 1.00 | 0.50 | 13.93 | 1.00 | 0.05 | 24.60 |
| **Test 2** | 147 | 2 | 7223 | 98 | 1.00 | 0.50 | 16.93 | 0.95 | 0.40 | 11.32 |
| **Test 3** | 221 | 3 | 7149 | 97 | 0.98 | 0.61 | 6.95 | 0.94 | 0.72 | 9.39 |
| **Test 4** | 295 | 4 | 7075 | 96 | 0.97 | 0.80 | 3.56 | 0.93 | 0.74 | 11.21 |
| **Test 5** | 737 | 10 | 6633 | 90 | 0.95 | 0.80 | 7.07 | 0.88 | 0.84 | 5.91 |
| **Test 6** | 1474 | 20 | 5896 | 80 | 0.92 | 0.84 | 7.16 | 0.87 | 0.85 | 6.58 |
| **Test 7** | 2211 | 30 | 5159 | 70 | 0.88 | 0.85 | 5.97 | 0.86 | 0.84 | 8.59 |
| **Test 8** | 2948 | 40 | 4422 | 60 | 0.88 | 0.86 | 4.59 | 0.86 | 0.85 | 3.69 |
| **Test 9** | 3685 | 50 | 3685 | 50 | 0.87 | 0.86 | 7.16 | 0.87 | 0.86 | 5.64 |
| **Test 10** | 4422 | 60 | 2948 | 40 | 0.86 | 0.86 | 7.33 | 0.85 | 0.84 | 11.09 |
| **Test 11** | 5159 | 70 | 2211 | 30 | 0.85 | 0.86 | 6.56 | 0.86 | 0.86 | 6.29 |
| **Test 12** | 5896 | 80 | 1474 | 20 | 0.86 | 0.86 | 6.45 | 0.86 | 0.84 | 4.95 |
| **Test 13** | 6633 | 90 | 737 | 10 | 0.86 | 0.85 | 6.51 | 0.86 | 0.85 | 2.58 |
| **Test 14** | 7370 | 100 | | | | | | 0.85 | | |

Table 1 shows experimental results.
During the test phase, the results show a determination coefficient for the Ozone:

- ranging from 0.05 to 0.86 for CRPS NN
- ranging from 0.50 to 0.86 for Cluster NN

In order to obtain the same value in term of determination coefficient by using the CRPS strategy at different percentage of input, we compare results within the above different approaches.
The performances of cluster analysis choices are always higher than conventional benchmark. In particular, our results show three different behaviours linked to information data.
The first is related to selection up to 1% (corresponding to the 74 pattern on 7370) of cluster. In this case, we have an equivalent performance about 3% (corresponding to the 221 pattern on 7370) of CRPS. The second is related to increase the selection by cluster methods from 4% to 10% and we obtained the same performance up to 10% of CRPS. At the end, beyond 30% of patterns selected by CA methods, we achieved equivalent performances of CRPS greater than 85% in term of $R^2$.
In term of CRPS NN, we consider different percentages of input patterns and we observe a rapid increase of performance after the 10% of data. If we use input patterns greater than 10% of data, $R^2$ is greater than 0.8. The NN performances decrease in meaningful way for lower percentages of input patterns. In fact, $R^2$ is 0.05 and 0.75 for the 1% and 4% of entire dataset respectively.
The use of CA as pattern selection increases NN performances in a very significant way. The NN training obtained by use of 1% (74 patterns given by centroids coordinates) and 3% by total data of cluster gives $R^2$ ranging from 0.50 to 0.61.
Moreover, in Table 7, we also calculated the percentages of negative ozone prediction obtained by CRPS NN and Cluster NN. In general, CRPS presents higher percentages than Cluster NN. When we considered test 1, Cluster NN produce about 14% of negative O3 concentration with average level of 14.9 µg/m³ and standard deviation 25.8 µg/m³, whereas CRPS produce 25% with the average level of 40.9 µg/m³ and standard deviation 52.0 µg/m³.
The same consideration are verified for all over tests.
The trend of measured ozone and the ozone reproduced by the NN models are given in Figure 5 to Figure 12, in which it is possible to observe that that Cluster NN converges much faster than conventional algorithms with compatible clustering quality. In particular, test 10 is the best result for Cluster NN, while test 11 is the best result for CRPS NN.
In this case, the determination coefficient of Cluster NN (test 10) and conventional NN (test 11) is equal in testing phase (Figure 9 and Figure 11).

**Figure 5: CLUSTER NN (test 1)**

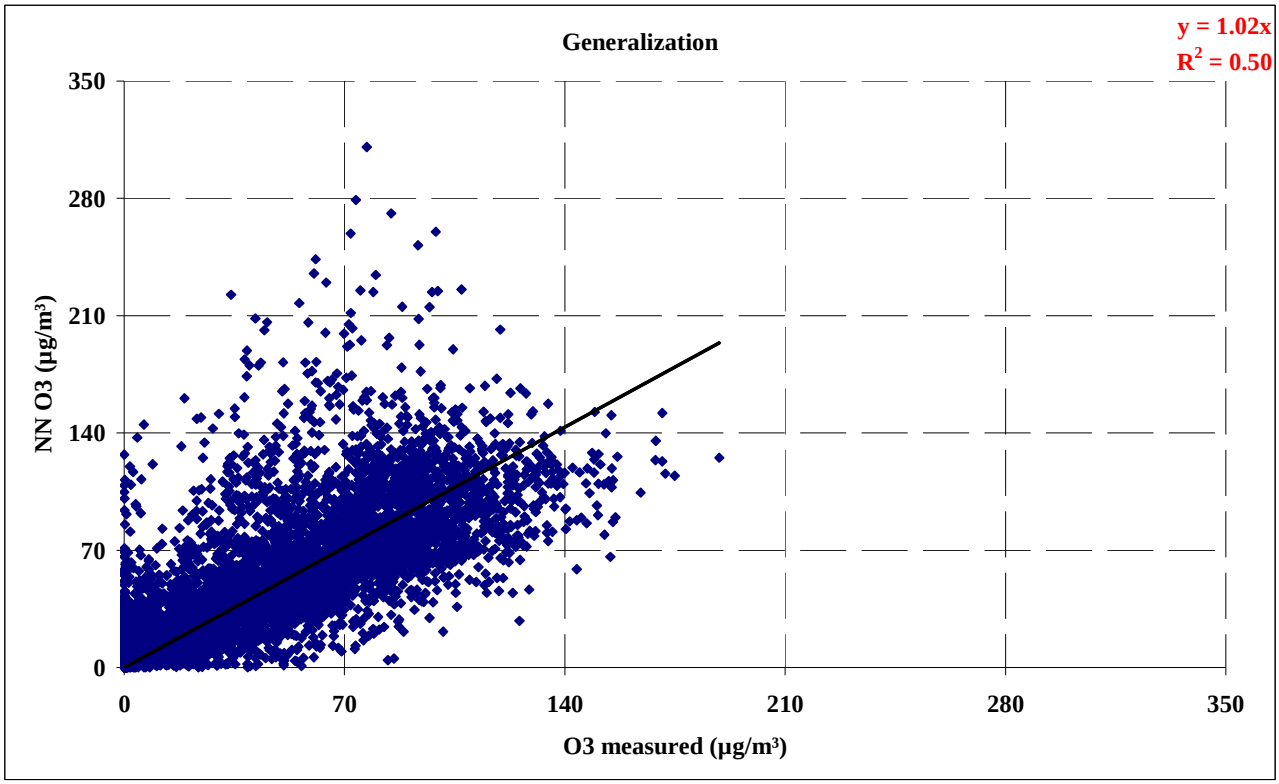


**Figure 6: Trend of Ozone as reproducing by Villa Ada monitoring station (test 1- Cluster NN)**

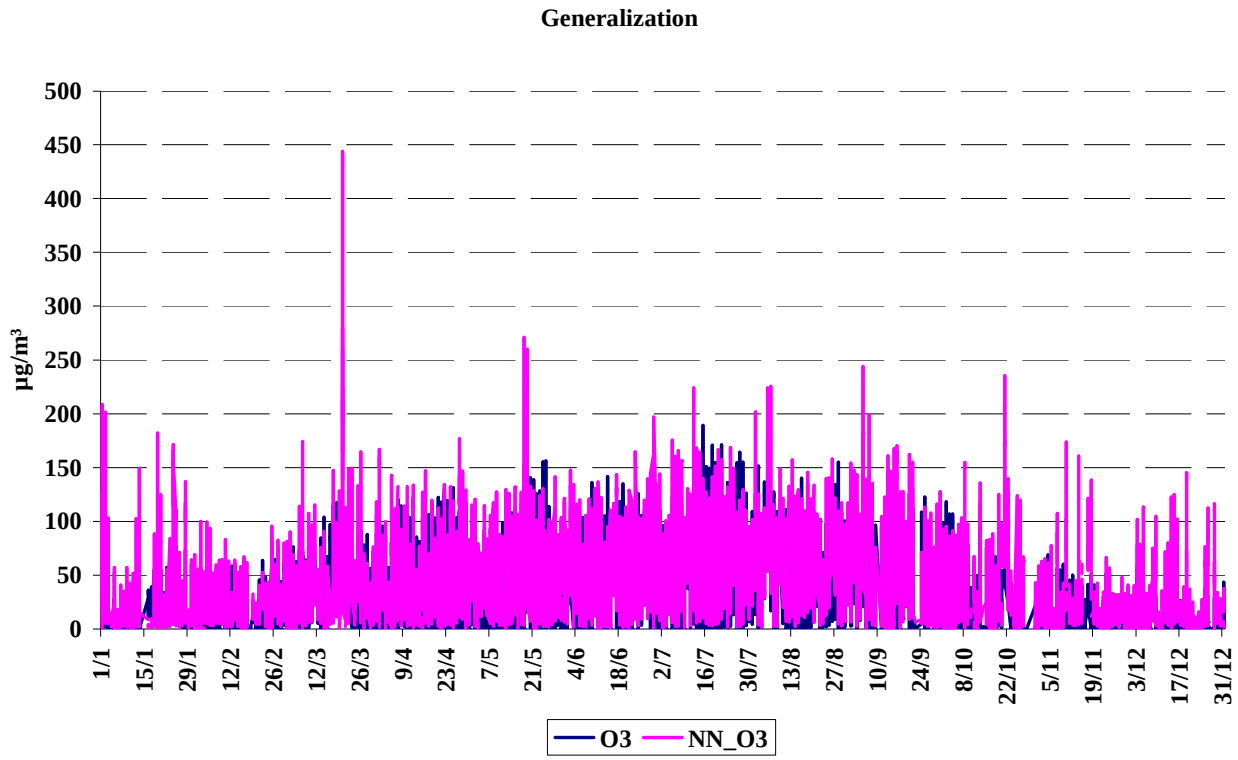



**Figure 7: CRPS NN (test 1)**

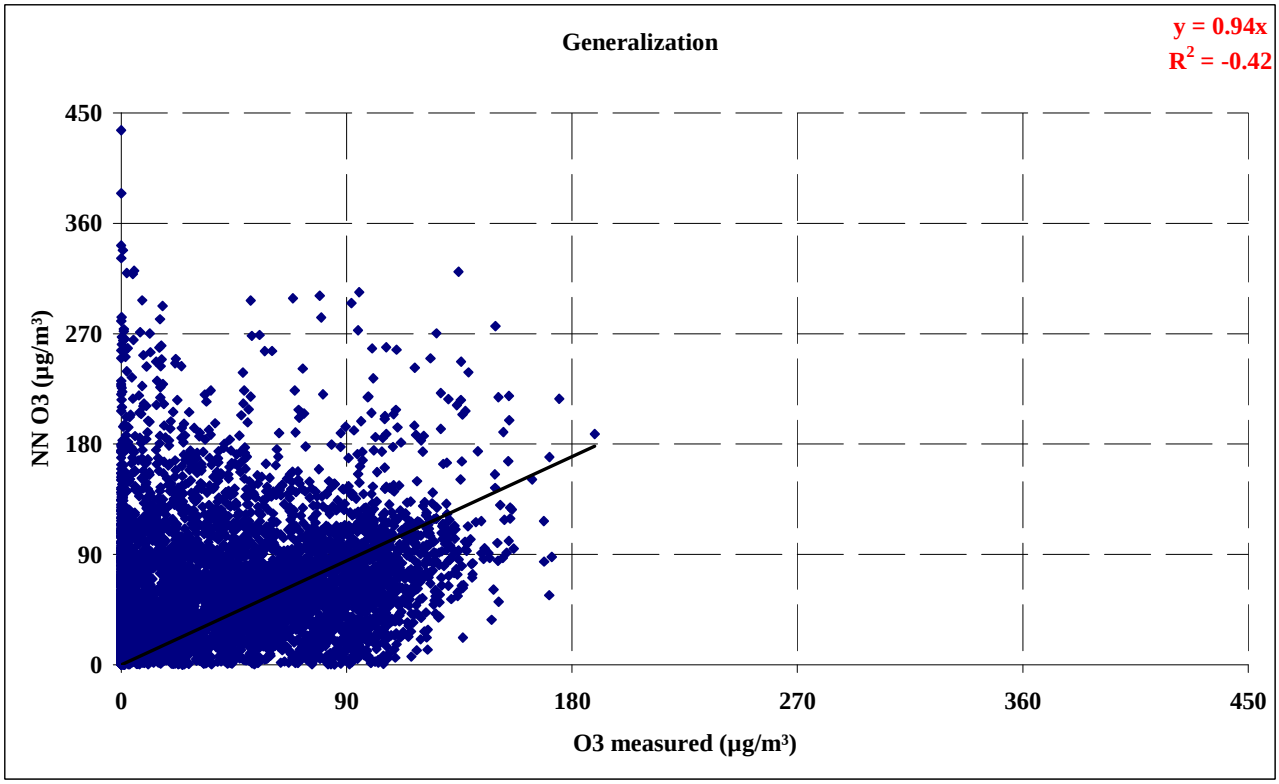


**Figure 8: Trend of Ozone as reproducing by Villa Ada monitoring station (test 1- CRPS NN)**

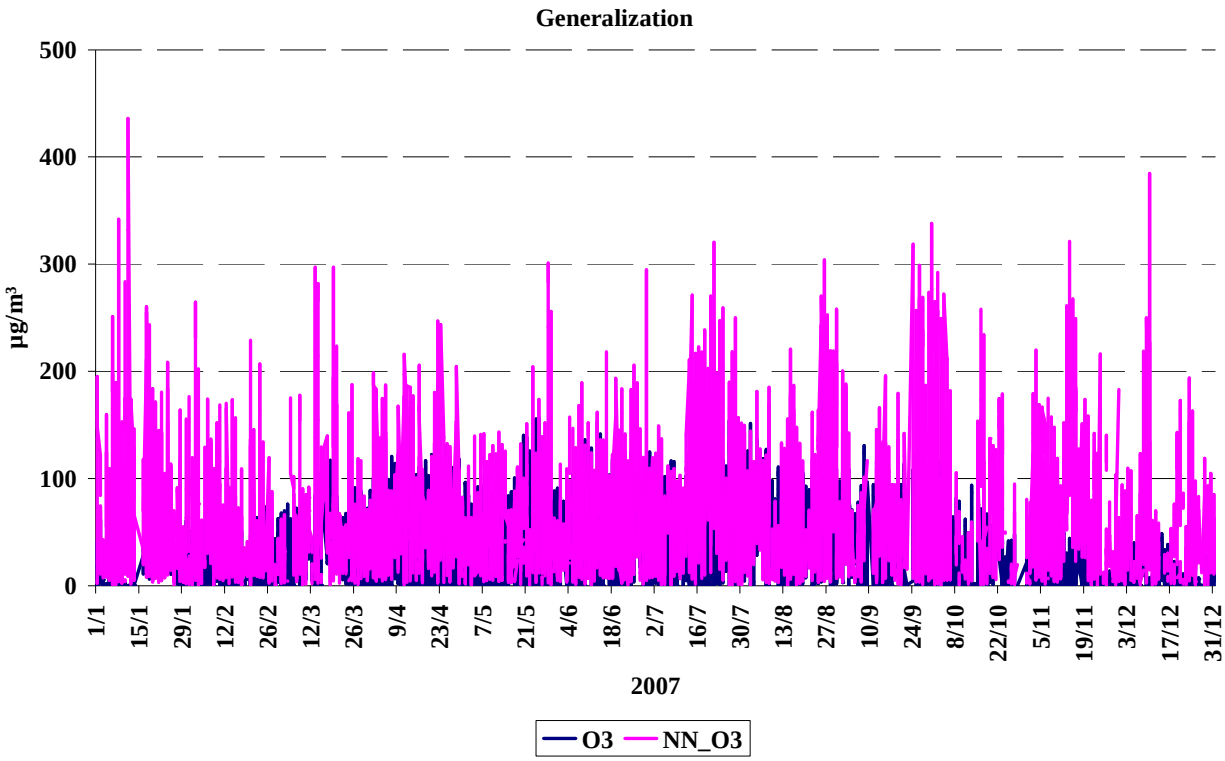



**Figure 9: CLUSTER NN (test 10)**

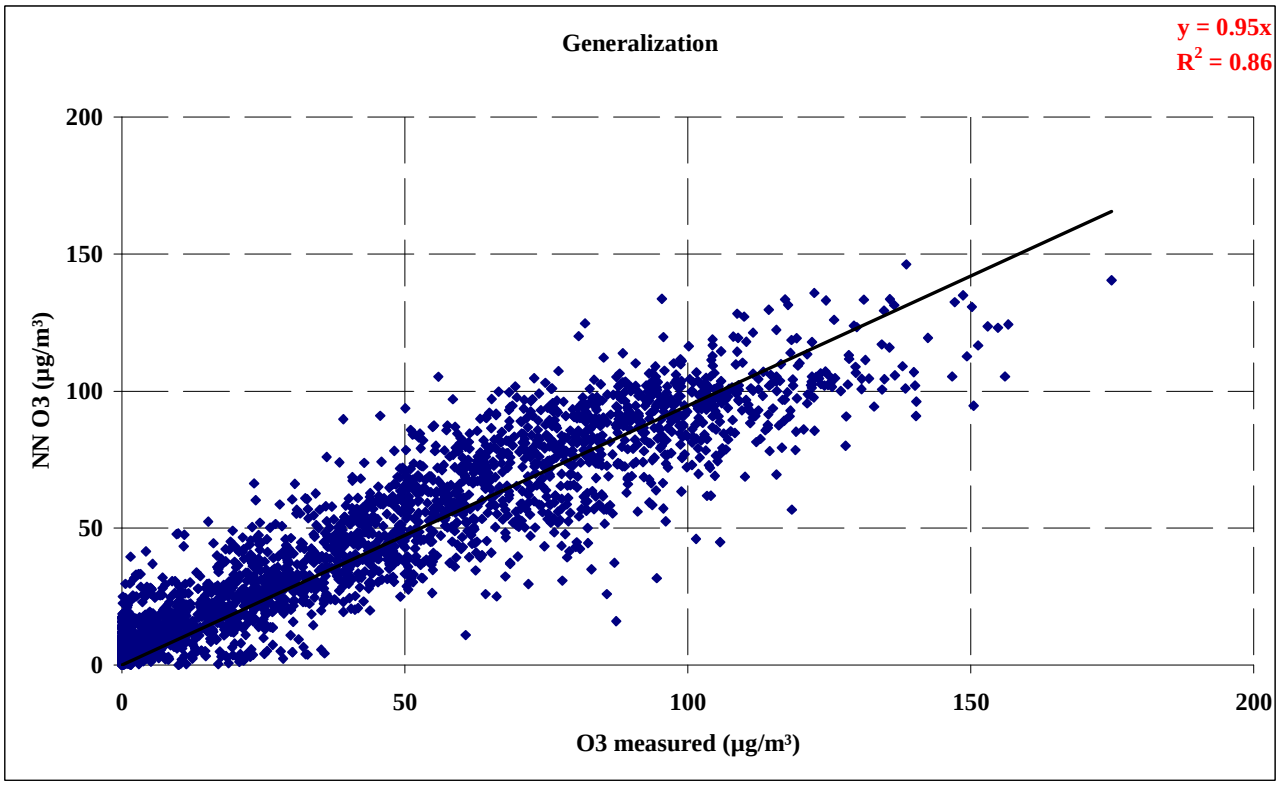


**Figure 10: Trend of Ozone as reproducing by Villa Ada monitoring station (test 10 - Cluster NN)**

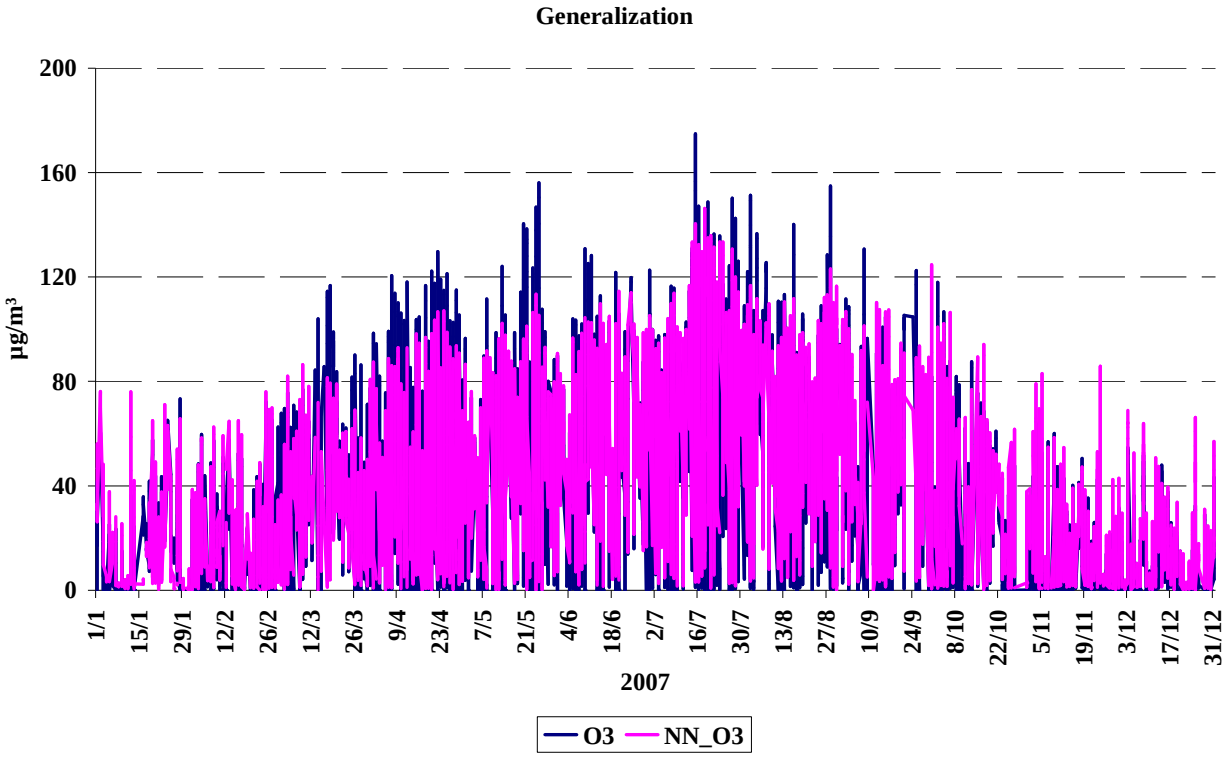



**Figure 11: CRPS NN (test 11)**

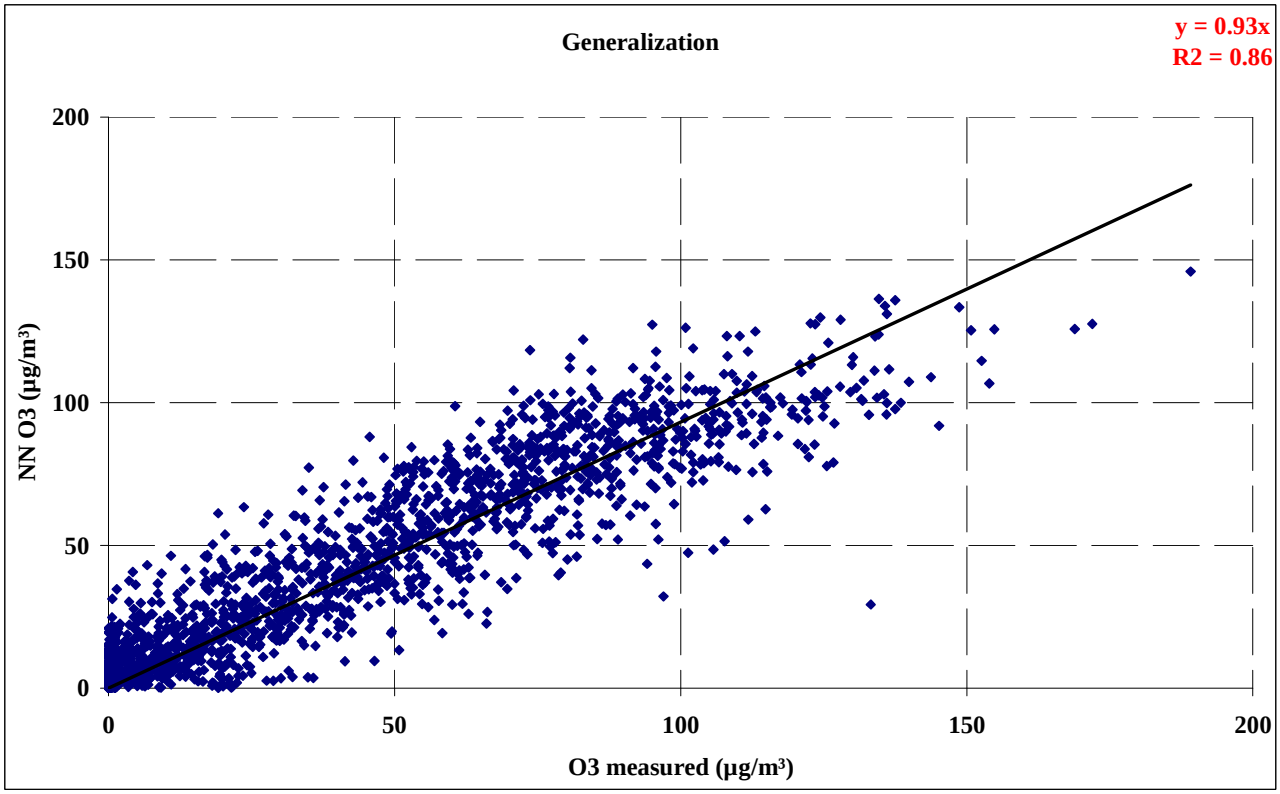


**Figure 12: Trend of Ozone as reproducing by Villa Ada monitoring station (test 11- CRPS NN)**

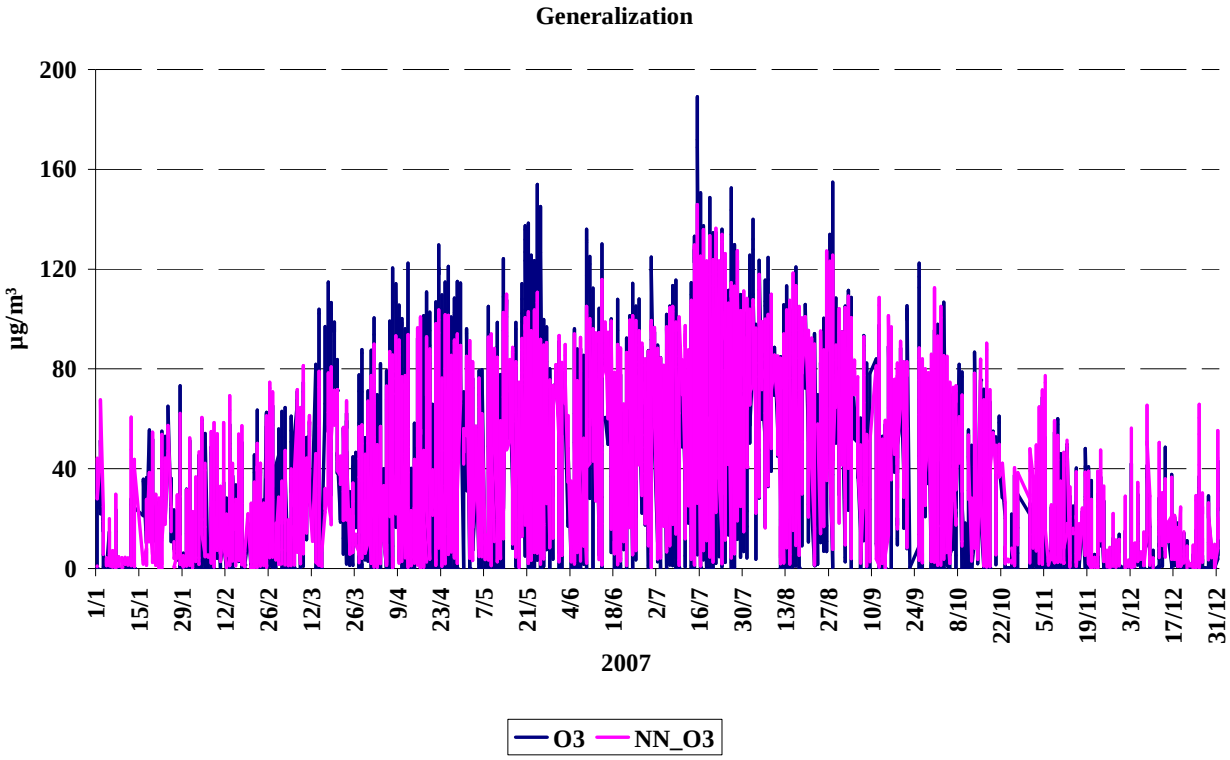
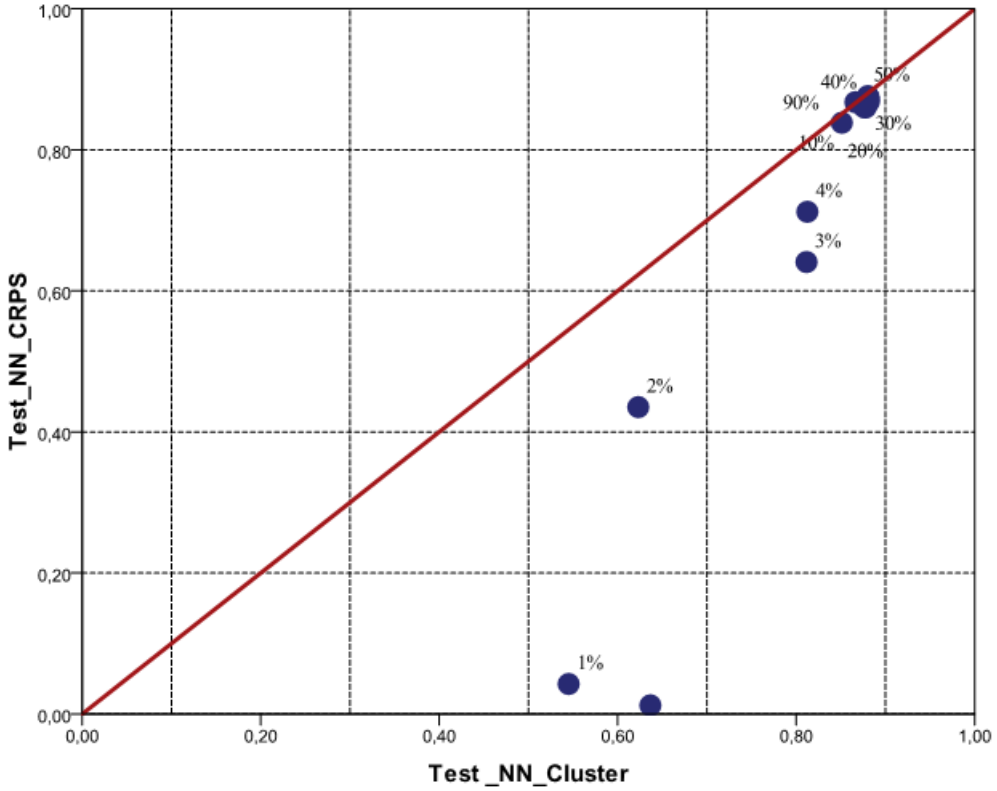



Figure 13 shows $R^2$ for different percentage of input patterns when using CRPS NN and Cluster NN. The same performances between the two model could verify if data are set along the diagonal (black line). The Cluster NN perform better than CRPS when we consider 30% of input pattern. In test 1 we obtain $R^2$=0.50 utilizing Cluster NN, whereas $R^2$=0.05 for CRPS NN. At 30% the performance of these two model are the same.

**Figure 13: Coefficient of determination in generalization phase**

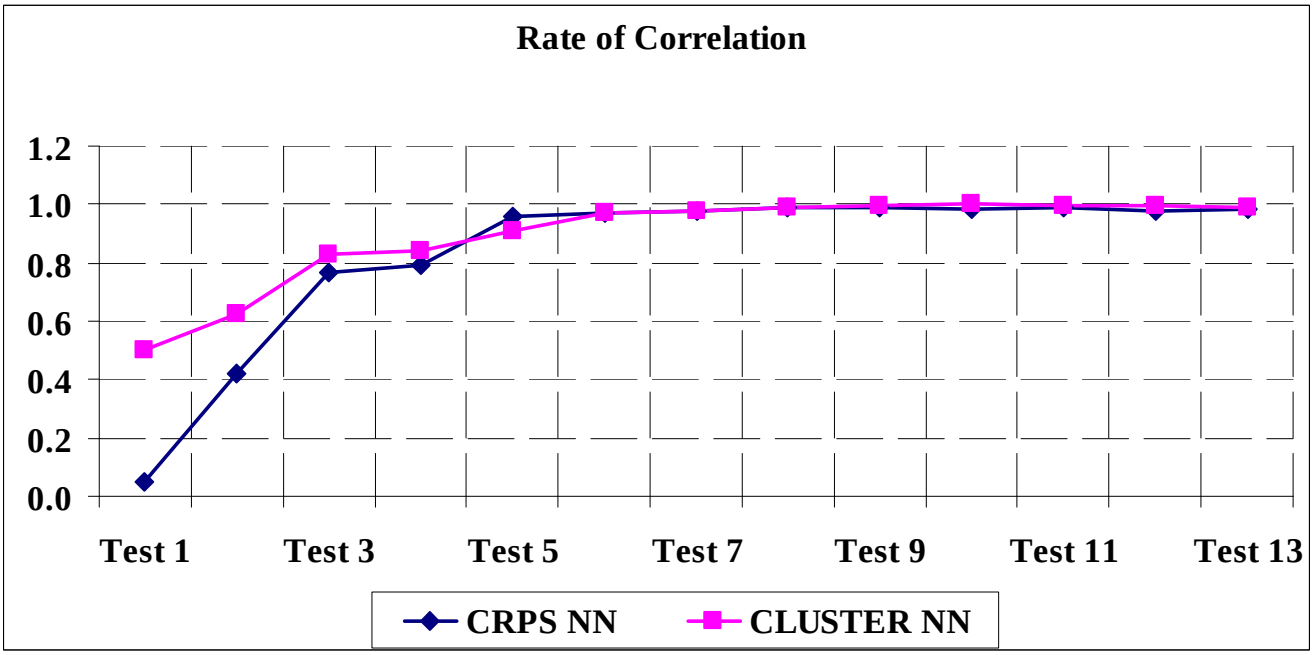


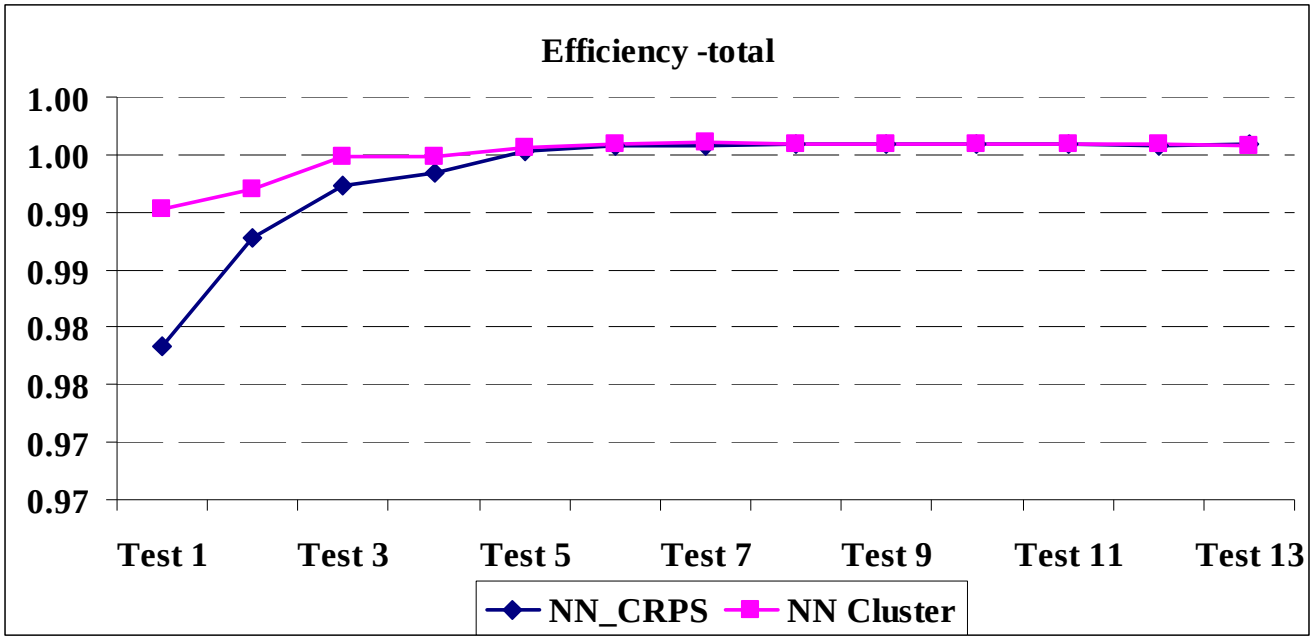

At the end, for measuring the efficiency of our NN, we consider the rate between the correlation coefficient calculated by generalization and by training phase (Rate Correlation-RC) and the rate of efficiency (see equation 3).

$$eff = \frac{F_0 - F}{F_0} \qquad (3)$$

We could observe many interesting results. Usually, the conventional way to train NN presents a marked maximum value of this RC corresponding to the well know decreasing of generalization performance of Neural Networks with the increase of percent of input training data (the so called over-fitting question). In particular, the Figure 14 shows the RC as calculated by our simulations. We observe that the maximum performance for the generalization is verified with 60% for Cluster NN and 70% for CRPS NN of input data for all trials. The rate of efficiency (Figure 14) shows that in the first tests Cluster NN outperforms CRPS NN, with values in the range of 99-100%.
This implies that, cluster analysis in the model contributes very much to a good prediction of ozone levels.

**Figure 14: Trend of Rate of Correlation at different simulations**




Our results are very encouraging and show that the NN model performance is improved using CA, as regards the conventional random pattern choice, in which randomly assign patterns based on relative number or percentage of cases.
Simulations based on cluster analysis show that NN converges much faster than conventional algorithms. Moreover, these results demonstrate that our approach is feasible and effective, resulting in a substantial reduction of data input requirement and outperform the other techniques applied in this contest. $R^2$ and rate of efficiency substantially show better performance in the combined forecast procedure.

## CONCLUSIONS

Our results show that the capability of the NN to capture the environmental information inside the data depended not only the learning methods used but also on the preliminary study of patterns, related to the quality of the data, used to train the network.
In particular, NN model performance for Ozone forecast is improved using CA, as regards the conventional random pattern choice, in which randomly assign patterns based on relative number or percentage of cases. This approach also gives the first recommendations for solving the patterns selection problem.
CA approach improves the classification accuracy and reduces the training time of neural networks significantly.
Cluster NN appeared to be satisfactory and revealed more subtle variations among patterns and an optimal AI technique for solving complex problems, not underestimated high-ozone episodes and over-estimated low-ozone events.
Simulations based on CA show that NN converges more rapidly than conventional algorithms. Moreover, these results demonstrate that our approach is feasible and effective, resulting in a substantial reduction of data input requirement and outperform other techniques applied in this contest and therefore, the combining techniques are more accurate than each individual methodologies and offer increasing performance as regards each method.

## REFERENCE


[1] Carter, W.P.L, 1990. “A detailed mechanism for the gas-phase atmospheric reactions of organic compounds”. Atmospheric Environment, 24A, 481-518.

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

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

[4] Gardner M.W, Dorling S.R., 1998. „Artificial Neural Networks (the Multilayer Perceptron)- E Review of applications in the atmospheric sciences”. Atmospheric Environment 32(14/15), 2627-2636.

[5] A.L. Dutot, J.Rynkiewicz, F.E. Steiner, 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.

[6] Nunnari, G., Nucifora, M. and Randieri, C., 1998. “The application of neural techniques to the modelling of time-series of atmospheric pollution data”. Ecological Modelling 111, 187-205.

[7] Monge Sanz, B. M. and Medrano Marques, N. J., 2004. “Total ozone time series analysis: a neural network model approach” Nonlinear Processes in Geophysics, 11, 683-689.

[8] D. Hawkins, 1980. “Identification of Outliers”. Chapman and Hall, London

[9] Rojas R., (1996): Neural Networks: a systematic introduction, Springer-Verlag, Berlin Heidelberg

[10] L. Kaufman and P. J. Rousseeu, 1989. “Finding Groups in Data”. John Wiley and Sons.

[11] Fausett, L., 1994. “Fundamentals of Neural Networks”. Architectures, Algorithms and Applications. Prentice Hall, Englewood Cliffs, NJ 07632.

[12] Bishop, C.M., 1995. “Neural Networks for Pattern Recognition”. Clarendon Press, Oxford.

[13] Ripley, B.D., 1996. “Pattern Recognition and Neural Networks”. Cambridge University Press.