H13-177

OPTIMIZATION OF OZONE BY NEURAL NET FORECASTING USING CLUSTER ANALYSIS

A. Pelliccioni¹, R. Cotroneo¹ and F. Pungi¹

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

Abstract: In the present work, we predicted the levels of ozone in the urban area of Rome using neural networks (NN). The data used in this paper covered a period of one year, from January through December 2007 and come from a background monitoring station.

The aim of this paper is the prediction of ozone 24 hours in advance using conventional meteorological variables (T, RH, WS and GSR), primary pollutants variables (CO, NO, NO2) and other variables related to the relationships between primary and secondary pollutants in order to take into account the effects of sources’ emissions and the inclusion of photochemical reactions.

In this work, we have adopted a neural network model (NN) to relate the input data with ozone measured 24 hours in advance. Our approach considers different performances related to two methodologies of patterns selection during training phase. We compared these two techniques: the first adopts random pattern selection and the second cluster patterns selection. For this second approach, the idea is to use the cluster analysis for patterns selection and then apply an artificial NN.

The simulations show different significant results both in terms of forecasting capabilities accuracy and in term of generalization capability to predict ozone concentrations 24 hours in advance.

In term of NN performances, we consider different percentages of input patterns and we observe a rapid increase of performance after the 10% of data. If we use random input patterns within 10%-60% of data, R² range from 0.55 up to 0.71. The use of cluster analysis as pattern selection increases NN performances in a very significant way. The NN training obtained by use of 10% and 60% by total data of gives R² ranging from 0.70 to 0.78.

Our results are very encouraging and show that NN model performance is improved using cluster analysis in respect of the conventional random pattern choice. In fact, simulations based on cluster analysis show that NN converges more rapidly and accuracy of the prediction is very accurate.

Key words: Ozone, artificial neural networks, Data mining, cluster analysis, k-means.

INTRODUCTION

Air quality problems produced by high levels of ozone (O3). Ozone is one of the most critical pollutants and causes health problems especially in large metropolitan areas, like Rome, where the emissions are relevant and could be the cause of greater exposure of the population [11] and damage to ecosystems. It becomes meaningful in regions during summer under unstable turbulence conditions [8] when relevant photochemical production due to precursors emitted during combustion of fossil fuels by industry and transportation can be observed.

Ozone is a reactive gas and presents concentrations which are dependent both from the meteorological conditions and seasonal effects especially in the Mediterranean area characterized by high level of temperature and solar radiation that cause higher ozone concentration. Often, it is difficult to determine the environmental effects of ozone contributions under different atmospheric conditions [2].

The forecasting of ozone levels is very complicated to obtain as described in different studies [1] [5]. For ozone models the most difficult problems to deal with are the simulation of chemical reactions that occur in lower levels of atmosphere, the contribution due to long range transport and turbulence conditions [3].

As known, one of more important statistical tools to forecast air pollution data is the neural network (NN) [9] that can be used to evaluate non linear behaviours.

The aim of this study is to develop an optimized NN model that forecast ozone concentrations one day in advance by using an ad hoc pattern selection.

The accuracy of the ozone forecast one day before is strongly dependent on the process of input patterns selection that is strategic to optimize the performance of NN model. In our work, we compared these two techniques: the first uses random pattern selection and the second uses cluster patterns selection. For this second approach, the idea is to use the cluster analysis for patterns selection and then apply an artificial NN.

1 DATASET DESCRIPTION

Our time series, related to the calendar years 2007, come from an urban background monitoring station of Rome (Villa Ada monitoring station). The city of Rome is characterized by frequent ozone peaks, associated with hot sunny days and instable-turbulence conditions. Other important factors derive from the main primary pollutants (NO, NO2, CO) coming from the principal urban sources. Villa Ada monitoring station is not directly affected by urban polluting sources inside the city and therefore represents a typical sub-urban situation with high ozone concentration levels located in the NNW direction.

This study used a conservative number (four) of pollutants (ozone and other relevant pollutants), conventional meteorological variables and other variables related to the relationships between primary and secondary pollutants in order to take into account the effects of emissions and photochemical reactions, in order to maintain parsimony and keep the resulting statistical models simple enough for meaningful comparison.

Our dataset concerns about 8760 hourly patterns surveyed every day in order to obtain the hourly distribution in the 24 hours and 11 variables:

1. observed pollutant variables:
   - Carbon monoxide (µgm⁻³) - CO -
   - Nitrogen Oxide (µgm⁻³) -NO -
   - Nitrogen Dioxide (µgm⁻³)-NO2 -
   - Ozone (µgm⁻³) - O3 - (Input/Output variable)
- CO/NOX that takes into account emission factors by traffic sources
- O3/NOX that represents dimensionless parameter linked to the photochemical production from precursors emitted primarily during combustion of fossil fuels by industry and transportation

2. meteorological variables:
- Temperature (°C) – T-
- Global Solar Radiation (Wm⁻²) – GSR-
- Relative Humidity (%) - RH -
- Wind Speed (m/s⁻¹) - WS -

Table shows general statistics related to the characteristics of pollutants and meteorological parameters used in this study. We analyzed the following statistical characteristics of the distribution of hourly data: mean, standard deviation, maximum and minimum values of ozone. We observed that the maximum hourly ozone for 2007 lies around a value of 189.1 µg/m³ during summer season (15/07/2007 h.15.00) whereas the maximum hourly of CO is about 4.1 µgm⁻³ during winter season(11/01/2007 h.23.00). We also observed a considerable variability in the time series (CV=102.3%). In our analysis missing values are not accounted. We also examined that uncertainties in ozone predictions are found to be most strongly correlated with uncertainties in the NO₂, wind speed, relative humidity. Moreover, global solar radiation was analysed to investigate the ozone correlation with photochemical reactions.

<table>
<thead>
<tr>
<th></th>
<th>CO (µg/m³)</th>
<th>NO (µg/m³)</th>
<th>NO₂ (µg/m³)</th>
<th>O₃ (µg/m³)</th>
<th>T (°C)</th>
<th>RH (%)</th>
<th>GSR (W/m²)</th>
<th>WS (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.61</td>
<td>22.32</td>
<td>43.99</td>
<td>36.62</td>
<td>12.97</td>
<td>73.23</td>
<td>125.42</td>
<td>0.78</td>
</tr>
<tr>
<td>SD</td>
<td>0.38</td>
<td>41.80</td>
<td>26.25</td>
<td>37.46</td>
<td>7.09</td>
<td>19.62</td>
<td>221.95</td>
<td>0.79</td>
</tr>
<tr>
<td>CV (%)</td>
<td>61.98</td>
<td>187.33</td>
<td>59.67</td>
<td>102.32</td>
<td>54.64</td>
<td>26.80</td>
<td>175.57</td>
<td>101.28</td>
</tr>
<tr>
<td>Min</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>10.0</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Max</td>
<td>4.1</td>
<td>398.7</td>
<td>156.8</td>
<td>189.1</td>
<td>37.0</td>
<td>97.0</td>
<td>1002.0</td>
<td>5.90</td>
</tr>
<tr>
<td>N</td>
<td>8260</td>
<td>8277</td>
<td>8277</td>
<td>8279</td>
<td>8738</td>
<td>8760</td>
<td>8760</td>
<td>8760</td>
</tr>
<tr>
<td>Missing</td>
<td>500</td>
<td>483</td>
<td>483</td>
<td>481</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: General statistics

Usually, the pollutants distribution presents a skewness and the identification of outlier [6] situations has effects on health. In our case, Ozone distribution is highly skew (Figure ). In fact, about 97% of patterns belonging to the class 0-120 µgm⁻³, while less than 0.1% are above the information threshold (180 µgm⁻³). In this context, the pattern selections could play a crucial role in order to simulate rare events.

2 METHODS

In our paper, we have adopted a neural network (NN) to relate the input data with ozone 24 hours in advance, considering two different methodologies of patterns selection. The first uses random pattern selection and the second uses cluster patterns selection. For this second approach, the idea is to use the cluster analysis (CA) as patterns selection technique, enhance the learning capabilities and reduce the computation intensity of neural network model by pre-processing data using k-mean algorithm. Random pattern selection consists of selecting a random subset of patterns selection into training so that the size of the set is reduced whereas its representativeness power is not affected. The complement data are used in the generalization phase for evaluating its performance. It is noted that the separation of the dataset into two groups was performed randomly so that each data could have equal chances to be picked.

The second approach utilises CA exclusively to select the best and significant patterns in order to choose the NN weights. CA was conducted by non hierarchical method, k-means technique that can be used to group a large number of patterns efficiently. The K-means technique [7] is unsupervised algorithms that solve the well known clustering problem and classify or group, based on some similarities, a given data set into K homogeneous clusters. The k-means algorithm also finds the centroid of a group of data sets. Centroids constitute the new dataset, that represent better the natural distribution of the training phase, applying to the NN 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.

![Figure 1. Ozone distribution](image)
phase. In fact, the centroids can minimize the mean-squared error of our original dataset. After the neural network has been successfully trained, its performances are tested on 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.

For the network training, we fix same percentages of input patterns from 1%, 3%, 10% up to 60% of the original data for each simulation leaving from 99% up to 40% respectively as the testing phase, to evaluate the performance of generalization of the model. We select these percentages using random pattern choice and CA technique.

As NN architecture we adopted the classical Multi Layer Perceptron (MLP) that is the most commonly used neural network in the field of air quality prediction [4]. As activation function, we used the conventional sigmoid that approximates nonlineairities. For selecting the best hidden neurons, during training phase we used the NN with 15, 20, 25 and 30 neurons for hidden layer.

At the end, we chose as NN architecture a single MLP with one hidden layer of 20 neurons, that gives the best performances in terms of minimizing the error function and computational efficiency, (Table 2). As algorithm, we adopted conjugate gradient that minimizes network output error and accelerates the convergence rate by searching optimal solutions. For this algorithm, the computation time is proportional to the number of weights selected.

Finally, the 3000 number of epochs has been determined after several experiments in order to avoid over-training.

Table 2. Neural Networks architecture

<table>
<thead>
<tr>
<th>NEURAL NETWORK MODEL</th>
<th>MLP 12-20-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIDDEN NEURONS</td>
<td>20</td>
</tr>
<tr>
<td>ALGORITHM</td>
<td>CONJUGATE GRADIENT</td>
</tr>
<tr>
<td>EPOCH</td>
<td>3000</td>
</tr>
<tr>
<td>ERROR FUNCTION</td>
<td>SUM OF SQUARE</td>
</tr>
<tr>
<td>HIDDEN ACTIVATION FUNCTION</td>
<td>LOGISTIC</td>
</tr>
<tr>
<td>OUTPUT ACTIVATION FUNCTION</td>
<td>IDENTITY</td>
</tr>
<tr>
<td>NETWORK RANDOMIZED</td>
<td>NORMAL</td>
</tr>
</tbody>
</table>

3 RESULTS AND DISCUSSION

The results of CA approach applied to NN (CANN) to forecast ozone one day in advance are compared with the Conventional Random Pattern Selection (CRPS NN), our benchmark, at different percentage of input patterns from 1% to 60% (from SIM1 to SIM8), excluding negative O3 concentrations predicted by NN during the generalizations phase.

Our work synthesises the results of 16 simulations (8 for each methodology) where each one is composed of a training phase and a testing phase.

Table shows the results of the generalisation phase coming from the final simulations related to different environmental performances indices for forecasting ozone one day in advance. As environmental performances indices, the following will be used:

- Coefficient of Determination (R²) that measures the global fit of the model.
- Mean Error (ME) that measures the forecast bias.
- Mean Percentage Error (MPE) that measures the average of percentage errors.
- Mean Absolute Error (MAE) that measures forecasts error in environmental time series. It is less sensitive to large forecast errors.
- Mean Absolute Percent Error (MAPE) that measures the accuracy and effectiveness of the model
- Relative Absolute Error (RAE) and Root Relative Square Error (RRSE), in which the errors are normalized.

The table 3 shows that CANN performed and predicted better than CRPS. In term of global fit, CANN performs better (R² from 0.34 to 0.76) than the classic CPRS (R² from 0.07 to 0.71). In particular, CANN perform better than CRPS when we consider SIM6 (40% of input pattern). Moreover, we observe that in SIM1 we obtain R²=0.17 utilizing CANN, whereas R²=0.07 for CRPS NN. This important result shows that CA technique is also efficient using small a amount of patterns during the training and, consequently, could be adapted to simulate rare events.

As regards the bias, the CANN models don’t decrease so much in respect of CRPS. In fact, we have 13.09 µg/m³ up to 8.68 µg/m³ at CANN simulations.

As regards the value of MAE, RRSE and RAE, we find a similar behaviour obtained with bias coefficient. Moreover, it has to be observed that MAE is higher than the bias ones.

In general, we observe that the best performances in terms of R², bias and MAE are obtained for SIM6 for CANN and SIM4 for CRPS.

Another aspect must be underlined. From a physics point of view, pollutants have to be positive and negative values have no sense.

Table 3 also shows the percentages of negative ozone prediction obtained by CRPS NN and Cluster NN, where negative predictive values are the proportion of ozone that are not correctly forecasted by the NN model. In general, CRPS NN presents higher percentages than CANN. When we considered SIM2, CANN produce about 22% of negative O3
concentration with average level of 19.32 µgm⁻³ and variation coefficient, that measures dispersion of a distribution, 98.98%, whereas CRPS produce 27% with the average level of 27.98 µgm⁻³ and variation coefficient 140.53%.

Table 4: Performance measures for prediction at different simulations

| TRAINING | GENERALIZATION - CANN | O3 <0 | | % | N | % | N | 7204 | BIAS µgm⁻³ | R² | ME µgm⁻³ | MPE | MAE µgm⁻³ | MAPE | RAE | RRSE | VC | MEAN µgm⁻³ |
|---------|-----------------------|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIM 1   | 1 | 77 | 99 | 13.09 | 0.34 | -7.70 | -190.60 | 29.52 | 13.09 | 0.91 | 1.33 | 30.26 | 91.80 | 25.53 |
| SIM 2   | 3 | 231 | 97 | 7058 | 13.69 | -3.83 | -278.20 | 25.07 | 13.69 | 0.77 | 0.90 | 22.04 | 98.98 | 19.32 |
| SIM 3   | 10 | 772 | 90 | 6549 | 7.52 | 0.70 | -0.04 | -163.34 | 15.32 | 7.52 | 0.47 | 0.29 | 5.85 | 117.95 | 5.37 |
| SIM 4   | 20 | 1545 | 80 | 5821 | 8.56 | 0.72 | -0.45 | -190.11 | 8.56 | 0.45 | 0.25 | 2.82 | 115.16 | 4.71 |
| SIM 5   | 30 | 2318 | 70 | 5093 | 8.53 | 0.74 | 0.06 | -185.61 | 8.53 | 0.43 | 0.22 | 1.85 | 103.23 | 4.08 |
| SIM 6   | 40 | 3090 | 60 | 4366 | 8.90 | 0.75 | 0.10 | -210.64 | 13.71 | 8.90 | 0.42 | 0.21 | 0.85 | 103.52 | 4.05 |
| SIM 7   | 50 | 3863 | 50 | 3638 | 9.60 | 0.55 | 0.68 | -177.19 | 18.15 | 9.60 | 0.56 | 0.45 | 8.60 | 154.41 | 5.69 |
| SIM 8   | 60 | 4636 | 40 | 2910 | 8.68 | 0.76 | 0.22 | -184.62 | 13.52 | 8.68 | 0.41 | 0.21 | 0.72 | 154.41 | 5.69 |

For measuring the efficiency of our NN, we also used two indices:

- DELTA defined as the difference between the CANN normalised rate of determination and CRPS NN normalised rate of determination (see equation 1)

  \[
  \text{DELTA} = \frac{(R_{CA}^2)^{\text{GEN}}}{{(R_{CA}^2)^{\text{TRAIN}}}} - \frac{(R_{CRPS}^2)^{\text{GEN}}}{{(R_{CRPS}^2)^{\text{TRAIN}}}}
  \]  (1)

  - the rate of efficiency defined as the difference between standard deviation by NN and root mean square error (see equation 2)

    \[
    \text{eff} = \frac{\sigma_{NN} - \text{RMSE}}{\sigma_{NN}}
    \]  (2)

In the Figure 2, we observe that the rate of efficiency and DELTA show that CA NN outperforms CRPS NN, in particular in the first simulation (1%). Usually, the conventional way to train NN presents a marked maximum value of DELTA corresponding to the well known decreasing of generalization performance of Neural Networks with the increase of percent of input training data. This implies that, cluster analysis in NN model contributes very much to a good prediction of ozone levels.

![Figure 2. Efficiency and variation of rate of determination](image-url)
It is important to note that for each appropriate pattern selection strategy, we succeed to reproduce in correct way of ozone one day in advance by using NN model. Our results confirm the capacity of NN to forecast ozone levels according the main research works [10] in an urban atmosphere. Moreover, the Neural Network is used to reproduce spatial correlation. This is a novelty on the pollutants prediction by NN, because, up to now, only the temporal trend was the target of the net. This work opens the way to the spatial-temporal reproduction by neural networks.

CONCLUSION

We have studied the problem of the best selection choice for NN to forecast O3 levels one day in advance.

The study shows that CANN model still provides higher performances and converges more rapidly than CRPS NN and in general conventional algorithms. CANN forecasting methodology is more reliable, feasible and effective, resulting in a substantial reduction of data input requirement and outperforming other techniques applied in this contest, especially in urban areas.

Moreover, our results demonstrate that the capability of the NN to capture the environmental information inside the data depended not only on the learning methods used, but also on the preliminary study of patterns, related to the quality of the data, used to train the network.

Our work suggests that using the cluster analysis technique as pattern selection improved the NN performances and suggested a way of optimizing and forecasting the environmental simulation using NN models approach, above of all during the rare events, when the higher pollutants levels are important to forecast.

This is the real novelty of our approach and gives the first recommendations for solving the patterns selection problem.

REFERENCES