Neural Networks and Co-Kriging techniques to Forecast Ozone Concentrations in Urban Areas

C. Carnevale
G. Finzi
E. Pisoni
V. Singh
M. Volta

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

https://scholarsarchive.byu.edu/iemssconference/2008/all/16
Neural Networks and Co-Kriging techniques to Forecast Ozone Concentrations in Urban Areas

C. Carnevale, G. Finzi, E. Pisoni*, V. Singh and M. Volta

DEA – University of Brescia, Via Branze 38, 25123, Brescia, Italy
({carneval, finzi, enrico.pisoni, vikas.singh, lvolta} @ ing.unibs.it)

Abstract: An integrated forecasting system, consisting of Neural Network (NNs) models and co-kriging techniques, has been developed to forecast maximum eight hours ozone (max8h) concentration, two days in advance, over an urban domain including Milan area of northern Italy. Total numbers of available measurement stations falling within the domain are 23. NNs perform the forecasting at each measurement location and the co-kriging algorithm interpolates the forecasting data all over the domain. NNs have been identified for the period of 2000-2006. Leave-One-Out Cross Validation (LOOCV) has been performed to validate the results of NNs. To perform spatial interpolation of the forecasted maximum daily eight hour (max8h) ozone, co-kriging has been used. For validation of the proposed forecasting system, 5 out of 23 stations have been selected. Year 2004 has been chosen as a test case year to perform the overall forecast. The validation results show good agreement in terms of statistical indexes. The proposed forecasting methodology represents a fast and reliable way to provide decision makers and general public with ozone forecasting data over an urban area.

Keywords: Ozone forecasting; Neural Networks; co-kriging, Spatial Interpolation.

1. INTRODUCTION

Since last decade tropospheric ozone episodes have become more and more critical over Europe, mainly in Southern metropolitan regions during summer months, due to the sun radiation significant role played in photochemical transformations of urban and industrial NOx and VOC emissions. Because of the environmental risk of ozone exposition, the EU Directive 2002/3/EC, following the WHO guidelines, prescribes air quality standards for ozone in terms of threshold values for health protection, population information and warning. In order to prevent critical episodes and to inform the population, proper real time alarm modelling systems have to be set up.

This paper proposes an integrated method, harmonizing Neural Networks [Corani (2005), Agirre-Basurko et al., (2006), Schlink et al., (2006), Sousa et al., (2007)] and co-kriging algorithms [Isaaks et al. (1990), Clayton et al. (1997)], to forecast the max8h ozone over an urban domain which includes Milan metropolitan area (Northern Italy). NNs are used to provide the forecast at the measurement locations, co-kriging to perform spatial interpolation of the forecasted data.

2. METHODOLOGY

Developed forecasting system consists of two parts:

1. Stochastic modeling system (Neural Network system) to forecast, two days in advance, daily max8h of ozone over each measurement station within the study domain.
2. Interpolation System to perform forecasted ozone maps over whole domain.

2.1 Neural Network system

Elman neural networks (Elman, 1990) have been identified to perform the forecasting for each measurement station. Elman NN implements a vectorial function $f: \mathbb{R}^Q \rightarrow \mathbb{R}^l$, where $Q$...
and $L$ are the dimension of the input and output vectors of the net respectively. The $l$-th element of the vector function $f$ for the $n$-th pattern is defined as:

$$f_l(y^n) = \log \text{sig}(\sum_{m=1}^{M}(OWL_{l,m} \cdot a^m_n) + g_l)$$

Where:

$$a^m_n = \tan \text{sig}(\sum_{q=1}^{Q}(IW_{m,q} \cdot v^n_q) + \sum_{w=1}^{W}(FW_{m,w} \cdot a^{n-1}_w) + b_m)$$

and $M$ is the number of the neurons in the hidden layer.

The matrices $IW (M \times Q)$, $OW (L \times M)$ and $FW (M \times M)$ are the input, output and feedback weight matrix respectively, and $b (M \times 1)$ and $g (L \times 1)$ vectors are the bias terms. NNs weights (IW, OW, FW, b and g) are tuned on a training dataset by means of a back-propagation algorithm (Mathworks, 2006).

Ozone measurement time series have been divided in identification and validation data sets; identification data set spans from 2000 to 2006 leaving out one year each time for validation.

2.2 Interpolation System

Interpolation System involves three steps:

1. Experimental semi-variogram and cross-variogram calculation;
2. Fitting and modeling of variograms using LCM (Linear Model of Coregionalization) [Pardo-Iguzquiza et al., 2002];
3. Estimation of max8h ozone over unmeasured locations using co-kriging method.

2.2.1 Experimental Semi-variogram and Cross-variogram Calculation

Variogram characterizes the spatial continuity or roughness of a data set and represents the variance of the increments. Empirical semivariogram $\gamma(h)$, computed as half of variogram, is a measure of the relation between pairs of points:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [z(u_{\alpha}) - z(u_{\alpha} + h)]^2$$

where $z(u_{\alpha})$, $z(u_{\alpha} + h)$ are the measurements in the points $u_{\alpha}$, $u_{\alpha} + h$; $N(h)$ is the number of pairs of points, the distance of which is $h$.

When primary data is not sufficient to represent the spatial structure, a more intensely sampled data, which is correlated with the primary data, is used to obtain the spatial structure of primary variable. Cross-variogram $\gamma_{zy}(h)$ represents the spatial relationship between primary $z(u_{\alpha})$ and secondary variable $y(u_{\alpha})$ and is defined as:

$$\gamma_{zy}(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [z(u_{\alpha}) - z(u_{\alpha} + h)] [y(u_{\alpha}) - y(u_{\alpha} + h)]$$

In present case, the primary variable is max8h ozone forecast at measurement locations over the domain and TCAM model [Carnevale et al., 2008] simulated ozone maps for summer 2003, calculated in the framework of an APAT (Italian Environmental Protection Agency, www.apat.gov.it/) project, has been used as secondary variable.

2.2.2 Fitting of Variograms and Linear Coregionalization Model

The empirical variograms are fitted by an analytical function (model). In this work the linear coregionalization model (LCM) [Pardo-Iguzquiza et al., 2002] has been applied. It is a sum of two or more proportional covariance (semi-variogram) models. A proportional covariance model is the simplest multivariate model used in geostatistics and is one in which all the semi-variograms $\gamma_j(h)$ are proportional to a single semi-variogram $\gamma(h)$ function:

$$\gamma_j(h) = b_j \gamma(h)$$
where, $b_{ij}$ are symmetric coefficients, that define a positive definite matrix $B=\{b_{ij}\}$.

Fitting a LCM comprises the following steps:

a) All direct semi-variograms and cross-semi-variograms are estimated for the same number of lags and the same lag distances $h$.

b) The number and types of elementary models and their ranges are postulated.

c) The sills (coregionalization matrix) are fitted by optimization technique (WSS [Pardo-Iguzquiza et al., 2002]).

2.2.3 Co-kriging

Co-kriging is an interpolation technique that allows one to better estimate primary variable if the distribution of a secondary variable is sampled more intensely than the primary variable. The co-kriging estimate is a linear combination of both primary and secondary data values and is given by

$$Z_0 = \sum_{i=1}^{n} a_i z_i + \sum_{j=1}^{m} b_j y_j$$

where $Z_0$ is the estimated max8h ozone at location $0$; $\{z_i\}_i^n$ are the forecasted, by NNs, max8h ozone at $n$ nearby locations and $\{y_j\}_j^m$ are the TCAM ozone at $m$ nearby locations;

![Figure 1. Block diagram of developed forecasting system](image-url)
Carnevale et al. /Neural network and co-kriging techniques to forecast ozone

\(a_{ij}^{m}\) and \(b_{ij}^{m}\) are the co-kriging weights. These weights are calculated by inverting covariance matrices of each distance pair and multiplying it with the covariance matrix of the distance pair from the estimation point. The covariance matrix are generated using the semi-variogram and cross-variogram models.

Fortran 77 GSLIB libraries [http://www.statios.com/, http://www.gslib.com/] have been used for co-kriging system. GAMV code has been used to calculate experimental Semi-variogram and Cross-variogram and COKB3D code has been used to calculate the max8h ozone estimates. For variogram fitting LCMFIT2 program [Pardo-Iguuzquiza et al., 2002] has been used. Block diagram of the developed system is shown in Figure 1, where \(\bar{z}_i(t-1)\) are the max8h ozone measurements for the previous day; \(\bar{x}_i(t)\) are the maximum first twelve hours ozone concentration for \(t\); \(T(t), T(t+1), T(t+2)\) are the mean temperature; \(z_i(t), z_i(t+1), z_i(t+2)\) are the NNs forecasted Max8h Ozone for each station and \(z_j(t), z_j(t+1), z_j(t+2)\) are the forecasted Max8h Ozone concentration all over the domain for the corresponding days.

The developed methodology has been tested over a densely inhabited and industrialized area located in the Northern Italy domain and including Milan.

3. RESULTS

3.1 Case study domain

Test case has been performed to forecast max8h ozone, for the entire year 2004, over Northern Italy domain, including Milan metropolitan area. The domain has a dimension of 60x60 km\(^2\), divided in 144 cells of 5x5 km\(^2\) (Figure 2). The domain has 23 measurement stations (Figure 2), and each measurement station has max8h ozone measured from 2000 to 2006. Test case and validation of the results have been performed in two steps. In the first step, NNs Leave-One-Out Cross Validation (LOOCV) has been performed to forecast max8h ozone and in second step, performance of whole system, including co-kriging system, has been tested and validated for 2004.

![Figure 2](image)

**Figure 2:** Forecasting domain in Northern Italy including Milan metropolitan area, showing all 23 measurement stations (stations in blue are validation stations).

3.2 Neural Network Validation

LOOCV has been performed for the period of 2000-2006 by leaving each year at a time. Excluded year is the validation data set and rest of the data set is training data set for the NNs. In this study, NNs inputs are the max8h ozone concentrations of previous day \((t-1)\), the maximum value of first 12 hour ozone for today \((t)\) and the forecasted daily mean temperature for \(t, t+1\) and \(t+2\). The target patterns are the maximum eight hours ozone for today \((t)\), tomorrow \((t+1)\) and day after tomorrow \((t+2)\). To perform each day forecast \(t, t+1\) and \(t+2\), three individual networks have been identified for each station.

Forecasted max8h ozone has been compared with the corresponding day of measurement. Correlation coefficient and mean error has been computed for each LOOCV year and for all
23 stations. Box plots of correlation coefficient and mean errors for \( t, t+1 \) and \( t+2 \) are shown in Figure 3 and 4 respectively. Year 2004 has been chosen as a test case year and detailed statistical indexes for selected stations (502, 525, 535, 536 and 542) are shown in table 1. It is clear that networks are able to correctly forecast max8h ozone but performance of the forecast get worse advancing in the day of forecast.

3.3 Forecasting system validation

In the second step, co-kriging has been performed using NNs forecasted max8h ozone values, excluding five stations (502, 525, 535, 536 and 542) out of total 23 stations. These stations have been used as validation stations for forecasting system. Spatially interpolated max8h ozone at these five stations has been compared with the max8h ozone measurement at the respective stations and statistics have been computed to validate the results. Computed statistics for five validation stations, for \( t, t+1 \) and \( t+2 \), are shown in Table 2.
Table 1: Statistics\(^1\) of the NNs validation for t, t+1 and t+2.

<table>
<thead>
<tr>
<th>Stations</th>
<th>502</th>
<th>525</th>
<th>535</th>
<th>536</th>
<th>542</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Err</td>
<td>0.94</td>
<td>1.51</td>
<td>1.63</td>
<td>0.86</td>
<td>-0.88</td>
</tr>
<tr>
<td>RMSE</td>
<td>13.29</td>
<td>15.73</td>
<td>13.60</td>
<td>14.72</td>
<td>12.52</td>
</tr>
<tr>
<td>Corr</td>
<td>0.94</td>
<td>0.92</td>
<td>0.96</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>SR</td>
<td>47%</td>
<td>47%</td>
<td>77%</td>
<td>64%</td>
<td>90%</td>
</tr>
<tr>
<td>SP</td>
<td>47%</td>
<td>45%</td>
<td>79%</td>
<td>62%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Table 2: Statistics\(^1\) of the forecasting system validation for t, t+1 and t+2.

<table>
<thead>
<tr>
<th>Stations</th>
<th>502</th>
<th>525</th>
<th>535</th>
<th>536</th>
<th>542</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Err</td>
<td>2.95</td>
<td>2.94</td>
<td>-3.66</td>
<td>2.25</td>
<td>2.57</td>
</tr>
<tr>
<td>RMSE</td>
<td>15.73</td>
<td>19.44</td>
<td>16.76</td>
<td>17.52</td>
<td>14.26</td>
</tr>
<tr>
<td>Corr</td>
<td>0.92</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>SR</td>
<td>57%</td>
<td>42%</td>
<td>77%</td>
<td>67%</td>
<td>87%</td>
</tr>
<tr>
<td>SP</td>
<td>53%</td>
<td>67%</td>
<td>52%</td>
<td>69%</td>
<td>77%</td>
</tr>
</tbody>
</table>

Figure 5: Comparison between measured, NN forecasted and co-kriged max8h ozone for time t, t+1 and t+2.

The performance of the forecast gets worse for the next days of forecast. For example, the correlation, which is close to 0.9 for t, decreases with time, and varies between 0.82-0.86

\(^1\) The indexes in Table 1 and Table 2: M. Err: Mean Error; RMSE: Root mean square error; Corr: Correlation; SR: Percentage of forecasted exceedances correctly predicted over measured exceedances (threshold of 120 μg/m\(^3\)); SP: Percentage of forecasted exceedances correctly predicted over forecasted exceedances (threshold of 120 μg/m\(^3\)).
for $t+1$ and 0.81-0.85 for $t+2$. If other statistics are considered, they also worsen with time. The degradation in the performance can also be marked while looking at SP and SR indexes computed by the forecasting system. It is clear that stations 535, 536 and 542 (located inside the Milan) give better performances than the stations 502 and 525, which are close to Milan highway.

Figure 5 shows the time series for 100 days of 2004, showing measured, NN forecasted and Co-kriged max8h ozone concentration ($\mu$g/m$^3$) for station 542 for $t$, $t+1$ and $t+2$. These figures also show the degradation in the forecasted results for next days forecast. Figures 6-8 show the spatial measured and forecasted images of max8h ozone concentration ($\mu$g/m$^3$) for 4th, 5th and 6th of August 2004. It can be seen that the developed methodology is able to reproduce spatial patterns of the daily max8h ozone over the domain.

Figure 3: Measured and forecasted maps of max8h ozone over the domain for $t$

Figure 4: Measured and forecasted maps of max8h ozone over the domain for $t+1$.

Figure 5: Measured and forecasted maps of max8h ozone over the domain for $t+2$. 
4. CONCLUSIONS

Developed methodology, integrating NNs and co-kriging is able to forecast daily maximum eight hours ozone three days in advance. The results of the methodology application show good agreement in terms of statistical indexes, i.e. showing fair correlation coefficient values. The quality of forecast gets worse for the next days of forecast. This methodology also reproduces the spatial patterns of the daily max8h ozone over the domain. The proposed forecasting methodology represents a fast and reliable way to provide decision makers and general public with ozone forecasting data on an urban area.

ACKNOWLEDGEMENTS

The research has been developed in the framework of the Pilot Project QUITSAT (Qualità dell’aria mediante l’Integrazione di misure da Terra, da Satellite e di modellistica chimica multifase e di Trasporto – contract I/035/06/0 – http://www.quitsat.it), sponsored and funded by the Italian Space Agency (ASI).

REFERENCES

Agirre-Basurko E., Ibarra-Berastegi G. and Madariaga I., Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area, Environmental Modelling & Software 21, 430-446, 2006.