Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 4 (2011) 95–102

Procedia Environmental Sciences

Urban Environmental Pollution 2010

Study on correlations between Lidar scattered light signal and air quality data in an industrial area

Juliana Steffens[a], Roberto Guardani[a,] *, Eduardo Landulfo[b], Paulo F. Moreira Jr.[a], Renata F. da Costa[b]

[a]*Chemical Engineering Department, University of São Paulo, Av. Luciano Gualberto, 380 Trav. 3, 05508-900 - São Paulo, SP, Brazil*
[b]*Nuclear and Energetic Research Institute, Av. Lineu Prestes 2242, 05508-000 - São Paulo, SP, Brazil*

Received date September 30, 2010; revised date January 30, 2011; accepted date January 30, 2011

Abstract

Results of a campaign on the use of Lidar in an industrial area in the city of Cubatão, Brazil, are presented. Scattered light signal at different heights was correlated with air quality data monitored at ground level using multivariate techniques, aimed at identifying in a quantitative basis the similarities in the behavior of groups of variables. By using neural networks as a non-linear association method, a clear correlation was obtained between the Lidar scattered light signal and hourly-averaged ground-level ozone concentration.

Keywords: Lidar; atmospheric pollution; industrial emissions; multivariate statistics; neural networksIntroduction

1. Introduction

Optical remote sensing techniques like Lidar (Light detection and ranging) have important characteristics concerning their application in atmospheric monitoring, since they are much simpler in construction, and enable real time detection of changes in optical response over large distances, even in hostile environments with large fluctuations of temperature and pressure. A Lidar is an active remote sensing instrument, meaning that it transmits electromagnetic radiation and measures the radiation that is scattered back to a receiver after interacting with various constituents of the atmosphere. Lidars use radiation in the ultraviolet, visible or infrared region of the electromagnetic spectrum. Different types of physical processes in the atmosphere are related to different types of light scattering. By selecting different types of scattering processes it is possible to measure atmospheric composition, temperature and wind. Active remote sensing is an important tool to study atmospheric processes as it offers many advantages over passive remote sensing systems. One of the main advantages is the high vertical resolution that can be achieved, since the small divergence of the laser beam defines measuring volumes of typically only a few cubic meters at ranges of tens of kilometers [1]. Lidar has special capabilities for remote sensing of different patterns of atmospheric variables, mainly in places of difficult access. Sophisticated Lidar remote sensing systems are nowadays being used to study many different aspects of the atmosphere and its components. This tendency has been widely reported in the literature.

This paper reports a study aimed at using a Lidar system for aerosol monitoring based on backscattered light as a virtual sensor to monitor air pollutant levels in the atmosphere. The correlation between backscattered light signal and air pollutant levels was obtained by fitting a neural network to the data. The Lidar system consists of a pulsed Nd:YAG laser emitting light at 532 nm, and a collecting system for the backscattered signal. The study consists of testing different correlations between the

* Corresponding author. Tel.: +55-11-3091-2277; fax: +55-11-3813-2380.
E-mail address: guardani@usp.br.

1878-0296 © 2011 Published by Elsevier Open access under CC BY-NC-ND license.
doi:10.1016/j.proenv.2011.03.012

Lidar signal at different vertical distances from ground, and air quality data collected from a ground-based monitoring station. Data characterization was based on multivariate statistical techniques (principal component and cluster analysis).

The study was carried out in the industrial area of Cubatão, in the Southeast of Brazil, located at the Atlantic cost, sea level, ca. 50 km from São Paulo, and one of the largest industrial sites in the country. In a region with ca. 40 km^2 there are 23 large industries, including a steel plant, an oil refinery, 7 fertilizer plants, a cement plant, and 11 chemical/petrochemical plants, adding up to 260 pollutant emission sources, besides the urban area, with ca 130 thousand inhabitants. A number of initiatives adopted since the 80´s have led to significant reductions in industrial emissions. However, the region still deserves much concern by authorities.

The environmental problems caused by the industrial activities are aggravated by the climate and topography of the site, unfavorable to pollutant dispersion. Cubatão is located in a narrow coastal plain surrounded by a steep mountain range to the north, west, and east, and by the sea to the south. At ca. 1 km west and northwest a 600-1000 m high mountain shell retains air circulation. Depending on meteorological conditions the atmospheric emissions by industries and local road traffic can accumulate, resulting in events of peak air pollution levels. During the day, winds blow from the sea to the continent, carrying pollutants to the mountains, where they are channeled into narrow valleys. Thermal inversions often occur during winter months [2]. Due to the local topography and proximity of the ocean, wind direction and velocity show daily changes that affect air quality, and frequent events of high pollutants concentration in the industrial area are recorded. In the last 6 years a significant number of events of high levels of PM_{10}, O_3, NO_2 and SO_2, the main air pollutants in the area, have been reported, frequently exceeding the legal air quality standards [2]. An illustrative map of the region is shown in Figure 1.

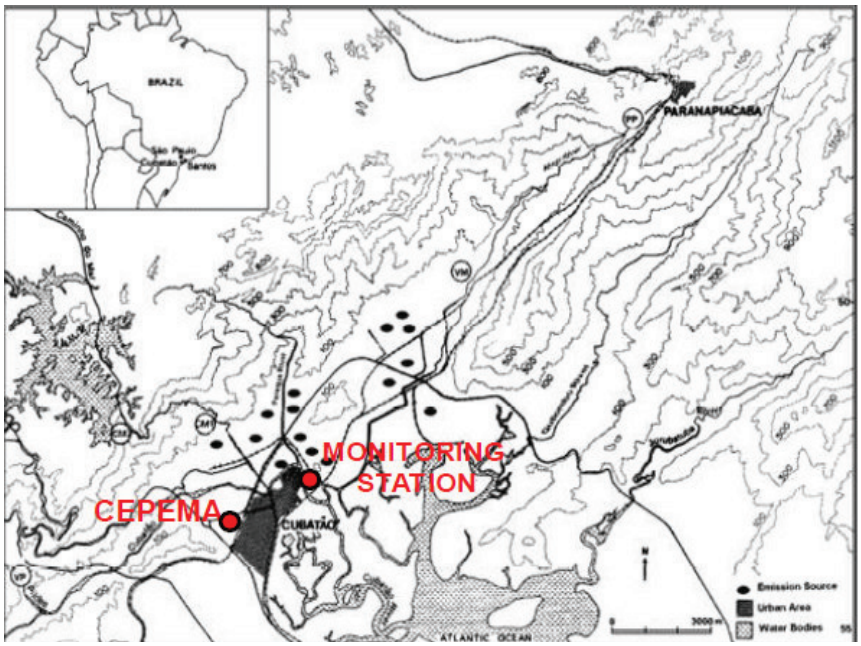

Figure 1. Map of the Cubatão area with indication of the monitoring site at CEPEMA, and the monitoring station at the city center (23°52'39'' S, 46°25'09'' W) [3].

2. Data collection

The main characteristics of the Lidar System are given in Table 1. This system was installed in the area of the Environmental Research Center (CEPEMA) of the University of São Paulo, located ca. 1 km west from the city center. During the campaigns, which took two months, October and December, 2009, the Lidar system operated during specific periods according to observed weather conditions (e.g., rainy, foggy, clear sky). The Lidar backscattered light signal at different altitudes was corrected for distance and air mass. Thus, the raw Lidar signal was converted into scattering ratio (ratio of the total aerosol plus molecular backscattering to the molecular backscattering).

Table 1. Specifications of the Lidar system (Commercial Raymetrics LR101-V-D200)

Emitter	
Laser type	Nd:YAG
Wavelength (nm)	532
Energy/pulse (mJ)	120
Pulse duration (ns)	6.7
Typical repetition rate (Hz)	20
Beam diameter (mm)	21
Beam divergence (mrad)	<0.17
Receiver	
Telescope type	Cassegrain
Primary mirror diameter (mm)	200
focal length (mm)	800
Detector	APD (at 1064 nm), PMT (otherwise)
Detection mode	Analogue, photon counting
Data acquisition	
Maximum spatial resolution (m)	7.5
Lidar Transient Recorder TR-20-160 (LICEL)	12 bit 20MHz analogue

Correlation of the Lidar information with air quality data was based on information provided by an automatic air quality monitoring station located in the center of the city, as shown in Figure 1. This station continuously monitors air quality and meteorological variables, and generates hourly averages of meteorological and air quality variables that are automatically transmitted and stored at the headquarters of the State Environmental Authority (CETESB) in São Paulo. The following meteorological variables are monitored: wind velocity ($m.s^{-1}$), wind direction (degrees from North), atmospheric pressure (mbar). The wind variables were transformed into the North to South (meridional) and East to West (zonal) components of wind velocity. The monitored air pollutants are: particulate matter (PM_{10},), SO_2, NO, NO_2, and O_3 (all in $\mu g.m^{-3}$). Prior to this study, specific monitoring campaigns have been carried out to compare hourly-averages of these variables at CEPEMA and at this monitoring station, indicating that the recorded values at both sites are strongly correlated, with an average correlation coefficient of 89.1 % for the monitored variables. A common data base was assembled, consisting of hourly averages of the Lidar signal at different heights from ground and the data from the monitoring station.

3. Data Processing

The first step in data processing involved the cleaning of the data base, mainly by eliminating missing data and gross errors (consisting of sensor failure, mainly). Then an exploratory study was carried out aiming at characterizing the data in terms of variability and correlations among variables or groups of variables. Based on this exploratory study the Lidar signal was grouped in four intervals of vertical distances from ground: up to 250 m, 250 to 500, 500 to 700, and 700 to 1000 m. The mean value of the scattering ratio in each interval was used in the study. These intervals were adopted because the aim was to study the correlation between the Lidar signal and air quality information at a local scale, as close to ground as possible, since the monitoring station monitors the variables at ground level. The lowest height was limited by the Lidar overlap distance, which is 180 m. The Lidar signal height was limited to 1000 m, which is ca. 200 m above the average height of the mountains that surround the region. The exploratory study consisted of examining the data characteristics by means of principal component analysis (PCA) and cluster analysis. The raw data base contained ca. 600 observations of each variable. The data have been checked manually for the presence of gross errors, or outliers. Numerical differences of variables were eliminated by working with standardized variables (zero mean, standard deviation equal to 1).

PCA consists of transforming the original variables of a multivariate system into non-correlated new variables (components) that are linear combinations of the original ones. Thus, from n original variables x_j ($j = 1,\ldots,n$) a smaller number of p non-correlated components e_i ($i = 1,\ldots,p$) are obtained, which are linear combinations of the original variables with the form of Equation 1:

$$e_i = w_{i1}x_1 + \ldots w_{ij}x_j + \ldots w_{in}x_n \quad (1)$$

in which the terms w_{ij} are the loadings, or weights, of variable x_j on the component e_i and are computed so that each component represents the maximum of the system variance in decreasing order. The technique is used to reduce the number of variables involved in an analysis, and to detect underlying relationships among groups of variables. The weights correspond to the eigenvectors of the covariance matrix of the original variables. Components are ordered according to the decreasing value of variances, represented by their eigenvalues. In order to eliminate numeric differences among variables, computations were based

on the correlation matrix. Result interpretation was based on the absolute value of the weights w_{ij}, [4]. A previous study based on PCA, made by Guardani et al. [5], aimed at obtaining information on the behavior of the main air pollutants in Cubatão (PM_{10}, O_3, NO_2 and SO_2) enabled the identification, based on quantitative statistical criteria, of the sceneries associated with high levels of air pollutants in the region.

Cluster analysis is a grouping technique based on common characteristics of subsets of variables or observations. A number of different techniques and criteria are commonly used. In this study, a hierarchical grouping technique was adopted to cluster variables, using the correlation coefficient as a distance measure, and different grouping criteria (single, complete, and centroid linkages). Clustering of observations was carried out according to the *k*-means method, based on the squared Euclidean distance between cluster centroids [6].

Fitting of three-layer feed-forward neural network models was based on the back-propagation algorithm, using in-house developed computer programs for model fitting and simulation. A detailed description of the model procedure applied to air quality data is presented in a previous publication by Guardani and Nascimento [7]. The fitting consisted of minimizing the quadratic deviation between computed and experimental values of the output variable, *E* (Equation 2).

$$E^{(m)} = \sum_{k=1}^{p} \left(y_k^{(m)} - O_k^{(m)} \right)^2 \tag{2}$$

where y_k and O_k are the experimental and calculated values of the output variable corresponding to the k^{th} observation in a total of p observations, calculated at the m^{th} presentation of observations to the neural network. All the collected observations were used in the model fitting process, in which the neural network configuration (number of neurons) and parameters of the learning process (number of presentations, learning rate) were varied. The sigmoidal function was adopted as the response for all neurons in the network. A total of 20,000 presentations of the data set to the neural network was adopted.

4. Results

Figure 2 shows an illustrative example of the Lidar data collected in the campaigns, corresponding to partially clear sky and occasional clouds at ca. 3.0 km altitude. Even under such dispersion-favorable conditions, higher light backscattering amplitude is observed at altitudes below ca. 1 km, indicating that the aerosol concentration remains significantly higher at low altitudes.

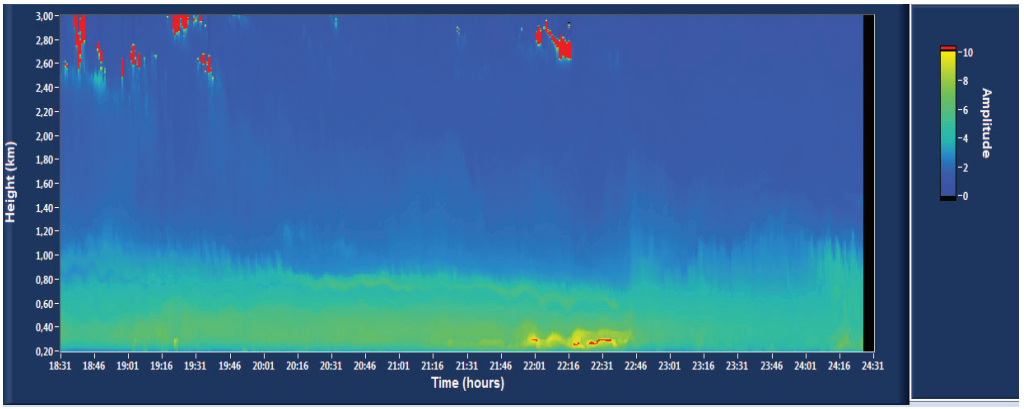

Figure 2. Illustration of Lidar backscattered signal amplitude at different heights from ground over time on November 6, 2009 (evening period).

Representative patterns of the Lidar data obtained in the campaigns are better visualized in terms of the scattering ratio (SR) plots shown in Figure 3, for two dates. In both cases, a condensation layer at ca. 1 km altitude is observed. Below 1 km, high aerosol concentration is observed, even under clear sky conditions (as in the Nov. 6 plot). The Nov. 18 plot shows higher aerosol concentration at higher altitudes in the morning period, followed by a rapid change in conditions after ca. 10h, due to increase in ground-level temperature and in wind velocity towards the monitoring site.

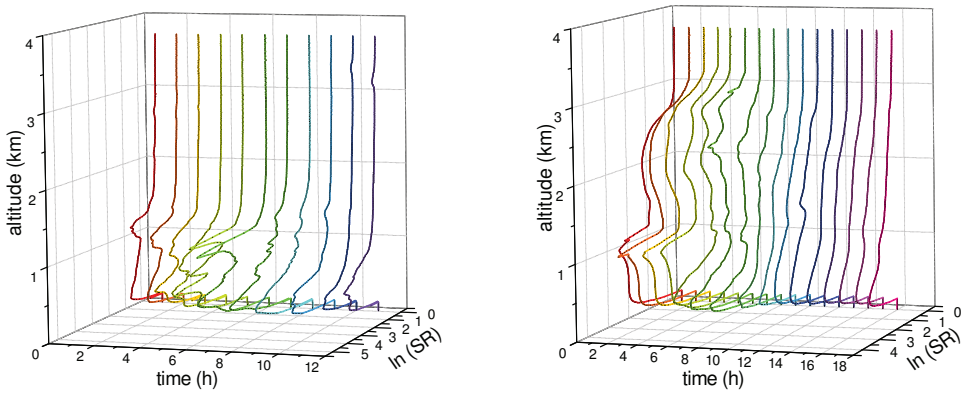

Figure 3. Lidar scattering ratio (SR) for two representative campaign days: November 1 and November 18, 2009.

Table 2 presents the correlation matrix for the variables included in the study. Expected correlations between nitrogen oxides and ozone are observed, and among the Lidar signal at different altitudes (last four columns/rows in the matrix), correlations are also observed between NO_2 and PM_{10}, and between the main pollutants and the North to South component of wind velocity (V, or meridional component). No significant correlation is observed between the Lidar signal at each height interval and the air pollutants (PM_{10}, SO_2, NO, NO_2, and O_3).

Table 2. Correlation matrix for the variables included in the study.

	Atm Press	**PM_{10}**	**NO_2**	**NO**	**O_3**	**SO_2**	**V***	**U***	**250m**	**500 m**	**700m**
PM_{10}	0.076										
NO_2	0.345	**0.543**									
NO	0.233	0.255	0.492								
O_3	-0.319	-0.143	**-0.552**	**-0.658**							
SO_2	-0.065	0.477	0.453	0.389	-0.208						
V*	0.051	0.394	**0.539**	0.346	**-0.601**	**0.561**					
U*	-0.089	-0.099	0.021	0.207	-0.377	0.051	0.172				
250m	0.264	0.322	0.270	0.235	-0.121	0.227	0.248	-0.064			
500 m	0.412	0.274	0.323	0.413	-0.341	0.275	0.280	-0.036	**0.828**		
700m	0.402	0.214	0.283	0.397	-0.312	0.193	0.222	-0.002	**0.817**	**0.968**	
1000m	0.396	0.129	0.243	0.352	-0.301	0.055	0.183	0.028	**0.740**	**0.880**	**0.950**

* **V**: North to South (meridional) component of wind velocity; **U**: East to West (zonal) component of wind velocity

PCA results are summarized in Table 3, which presents the weights of each variable (eigenvectors of the correlation matrix) on the first four components, which correspond to eigenvalues larger than 1. These four components represent more than 80% of the total variance of the data. The largest weight in each component is marked in bold. The largest weight in component 1 (PC1) corresponds to the Lidar signal at 500m, which is strongly correlated with the other Lidar signals. In component 2, the V component (North to South) of wind velocity is the most important variable, and is strongly correlated with SO_2. Thus, this pollutant is associated with wind coming from the industrial site. In PC3 the U component (East to West) of wind velocity is the most important variable, which is negatively correlated with ozone and PM_{10}. Thus, wind coming from the west (industrial site) is associated with both pollutants. PC4 is dominated by the atmospheric pressure, but this component corresponds to 8.8% of the total variance of the data, only.

Table 3. PCA Results: Component variances and weights of variables.

Component	PC1	PC2	PC3	PC4
% of Variance	40.8	18.6	12.0	8.8
Atm Press	0.211	-0.149	-0.194	**-0.680**
PM_{10}	0.214	0.235	0.487	-0.078
NO_2	0.291	0.322	0.102	-0.348
NO	0.290	0.227	-0.239	-0.057
O_3	-0.274	-0.304	0.442	0.125
SO_2	0.210	0.350	0.351	0.253
V*	0.254	**0.387**	0.058	0.141
U*	0.040	0.212	**-0.536**	0.473
250m	0.341	-0.280	0.189	0.209
500m	**0.399**	-0.264	0.031	0.109
700m	0.389	-0.310	-0.022	0.130
1000m	0.359	-0.337	-0.104	0.107

* **V**: North to South (meridional) component of wind velocity;
U: East to West (zonal) component of wind velocity

This behavior is more clearly visualized by means of the cluster analysis for grouping variables, as shown in the dendrogram in Figure 4. These results refer to the grouping of variables according to the average linkage criterion, using the correlation coefficient as a similarity measure. Similar dendrograms were obtained when the grouping was based on other hierarchical criteria, like single, complete, centroid linkages, and Ward method [6]. Two distinct clusters are observed: the first one contains the Lidar signal at different altitudes, and the atmospheric pressure. This is expected, since a clear correlation is observed among these variables (Table 2). Except for ozone, the main pollutants in the region and the V component of wind velocity form a second cluster, with a high similarity for this variable and SO_2. Ozone and the U component of the wind velocity show a distinct behavior from the other variables.

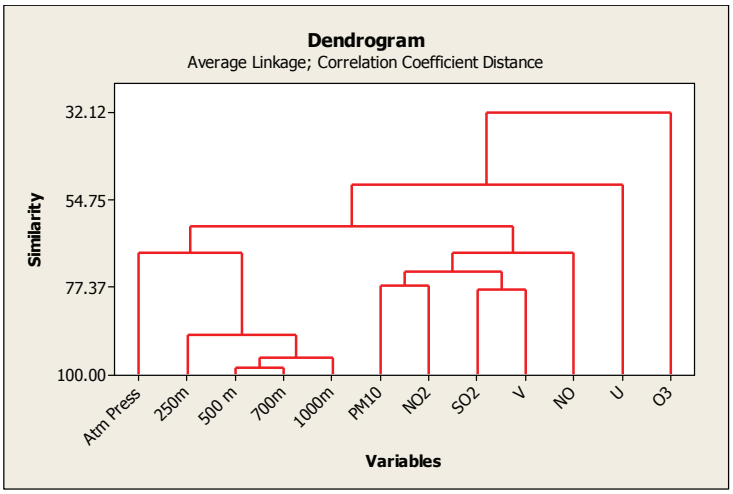

Figure 4. Dendrogram of the cluster analysis for grouping variables.

Clustering of observations was carried out in order to evaluate the ability to detect distinct patterns of the variables under different conditions of air quality and meteorology. In this case a non-hierarchical technique, namely the “K-means” method [6], was adopted. Clustering was carried out separately for the air pollutants and for the Lidar signal. Since, in this case, it is necessary to adopt the number of clusters, different numbers of clusters were tested, and the most coherent configuration in both cases resulted in three clusters. The results are presented in Table 4, each column consisting of the distance vector between the

centroid of each cluster and the grand centroid of all observations. In this case the grand centroid is kept equal to zero since each variable was standardized, i.e., centered at the mean value and scaled as multiple of the standard deviation. This procedure is adopted in order to eliminate numerical differences among different variables and to emphasize the variability of the data [6]. For the air pollutants, Cluster 1 has all pollutants located above the grand centroid (null vector), except for ozone, which is below average. Cluster 2 corresponds to observations with low levels for all pollutants included in the study. Cluster 3 contains observations with PM_{10} close to average, ozone above average, and the other pollutants below average. Concerning the Lidar signal at different altitudes, grouping of observations into three clusters resulted in the most coherent distribution, too. As shown in the table, Clusters 1, 2 and 3 show distances of the signals respectively above, around and below the grand centroid. Thus, although the range of values of the air pollutants is limited to the observed ones during the campaigns, distinct patterns of distribution of the variables were detected according to a quantitative criterion, i.e., the squared Euclidean distance between cluster centroids.

Table 4. Results of the k-means cluster analysis for the observations. Values are squared Euclidean distances between each variable in each cluster and the grand centroid (null column).

Variable	Cluster 1	Cluster 2	Cluster 3	Grand centroid
PM_{10}	0.845	-0.885	0.046	0
NO_2	1.147	-0.483	-0.442	0
NO	1.117	-0.182	-0.633	0
O_3	-0.919	-0.414	0.916	0
SO_2	0.800	-0.655	-0.085	0

Variable	Cluster 1	Cluster 2	Cluster 3	Grand centroid
250m	0.912	0.120	-1.652	0
500m	1.189	-0.139	-1.293	0
700m	1.227	-0.123	-1.394	0
1000m	1.202	-0.125	-1.354	0

The results of the cluster analysis indicate that PM_{10}, SO_2, NO, and NO_2 show similarities in behavior, and that higher-than-average levels of these pollutants are more likely to be observed when wind is blowing from North. Such patterns correspond to higher-than-average levels of the four Lidar signals used in this study.

As shown in the dendrogram in Figure 4, ozone shows a distinct behavior than the other variables included in the study, since it remained separated from the clusters that have been formed. Based on the results of the grouping of observations, shown in Table 4, higher-than-average levels of ozone are likely to occur when the other pollutants and the Lidar signals have lower-than-average levels. This observed correspondence between ozone levels and Lidar signal was verified by fitting a neural network model to the data. The objective was to verify the possibility of associating the Lidar signal patterns with ozone levels by means of non-linear relationships and the distribution of information among different neurons, characteristic of feed-forward neural networks. Different sets of the input variables were and different neural network configurations were tested. The best results were obtained with the following input variables (hourly averages): the four Lidar signals plus the meteorological information (atmospheric pressure, V and U wind velocity components), thus totalizing 7 inputs. The fitting algorithm was used to fit neural network models by changing the number of neurons in the hidden (intermediate) layer between 6 and 14, as well as learning parameters. Figure 5 shows the fitting results for a neural network containing 6 neurons in the hidden layer, in terms of the computed *versus* observed values of ozone concentration. As shown in the plot, a good agreement between computed and observed ozone levels has been achieved, with a coefficient of determination (R^2) equal to 0.928, and a trendline with slope close to 1 (0.913), and low bias (-1.096). Although limited to the range of ozone levels observed in the campaigns, this clear correlation is an indication of further possibilities for using information of the same kind as a means to estimate ozone levels based on the monitored variables of scattered light and wind velocity and direction.

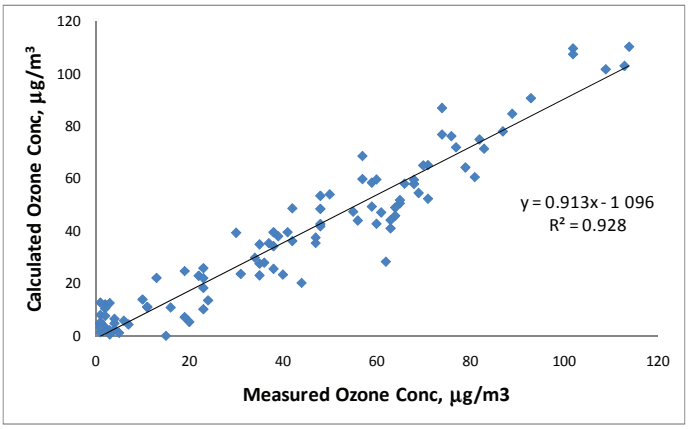

Figure 5. Comparison between neural network-computed and observed ozone concentration (hourly averages), for a neural network with 7 inputs (Lidar signal at 4 altitudes, atmospheric pressure, V and U wind velocity components), 1 output (ozone concentration) and 6 neurons in the hidden layer.

5. Conclusions

The reported results are the first ones obtained with a Lidar system for aerosol monitoring in the region of Cubatão. By associating the Lidar signal (scattering ratio) at different altitudes with the air quality data generated by a conventional monitoring station at ground level, important characteristics of the variables have been detected. In this sense, the application of adequate multivariate techniques has resulted in the identification of correlations between the monitored variables, individually and in groups, and has led to indications of patterns of distributions of the variables in situations corresponding to higher-than-average and lower-than-average levels of the main air pollutants in the region of Cubatão.

A particularly important result is the possibility of using the same information directed originally to the monitoring of aerosol levels as a means to estimate ozone levels, by associating the selected variables in a neural network model. The ability of the neural network to associate the selected input variables with ozone levels can be a consequence of the behavior detected in the cluster analysis, as shown in Figure 4 and Table 4, since ozone tends to behave differently than the other pollutants studied, and similarly to the wind velocity and direction. Although the observed ozone levels during the campaign period did not reach higher values, thus limiting the validity range of the test, the results bring an important perspective of using the Lidar signal for backscattered light, plus meteorological information on wind and atmospheric pressure, as a tool to estimate ozone levels.

Since this study is part of a program to install a mobile scanning Lidar system to map air quality in the Cubatão region, a more complete study, involving a substantially larger number of observations is presently being carried out, in order to confirm and to improve the indicative results reported in this text.

Acknowledgements

The authors kindly acknowledge the financial support by the Brazilian agencies CNPq and FAPESP, through INCT CEPEMA – National Institute of Science and Technology, Environmental Research Center, University of São Paulo, and by Petrobras.

References

1. G.L. Stephens, Remote Sensing of the Lower Troposphere. An Introduction, New York, 1994.
2. CETESB Companhia de Tecnologia de Saneamento Ambiental (Environmental Authority of São Paulo State), Relatório de Qualidade do Ar no Estado de São Paulo – 2007 (2007 Air Quality Report), CETESB, São Paulo, 2008.
3. M. Domingos, A. Klumpp, G. Klumpp. Air pollution impact on the Atlantic forest the Cubatão region, SP, Brazil, Ciên. & Cult. 50(4), (1998) 230-236.
4. I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
5. M. L. G. Guardani, M.H.H.R.B. Martins, R. Toyota, L.G. Morita, R. Guardani, 2009. Air quality data mining using multivariate statistical techniques: application to historical data from Cubatão, Brazil. In: 7th International Conference on Air Quality, 2009, Istanbul, 2009.
6. R. A. Johnson and D.W. Wichern, D. W., Applied Multivariate Statistical Analysis, Prentice-Hall, Upper Saddle River, 2002.
7. R. Guardani, C.A.O. Nascimento, Neural network-based study for predicting ground-level ozone concentration in large urban areas, applied to the São Paulo metropolitan area, Int. J. Environ. Poll., 22(4) (2004) 441.