# Tropospheric Ozone Prediction in Mexico City

Margarita Garfias Vázquez, Javier Audry Sánchez,* and Francisco Javier Garfias y Ayala

Facultad de Química, Universidad Nacional Autónoma de México
Circuito Exterior, Ciudad Universitaria, Coyoacán 14510. México, D.F.

**Abstract.** Two techniques are applied to forecast time series for the hourly ozone measured at Pedregal's recording station of the Automatic Network for Environmental Monitoring (RAMA for its acronym in Spanish) located in the metropolitan area of Mexico City. The techniques have been widely applied since last century: the autoregressive (AR) and the method of delays in an embedded space. The predicted values by the autoregressive method are somewhat less precise than those forecasted by the embedded space method, as presented below. It is intended to predict the maximum ozone daily concentration in advance to be able to alert the citizenship or for taking appropriate control measures. In this presentation, the models have as main limitation to be based only on the ozone time series; more robust models should take into consideration meteorological variables to increase precision. If it is roughly considered only the series formed by the hourly ozone series from January to May 1999; the series of daily ozone maxima have an standard deviation of around 0.05 ppm of ozone. In the most precise —the embedded space method— shown at the end of the article, the error standard deviation between predicted and real maximum daily ozone data is around 0.027 ppm of ozone, which shortens the gap, considering the total of the ozone maximum as a normal distribution.

**Key words:** Ozone measurements, Mexico City, autoregresive method, embedded space method.

**Resumen.** Se aplican dos técnicas para el pronóstico de series en el tiempo para el caso de la serie formada por las lecturas de ozono horario en la estación de Pedregal de la Red Automática de Monitoreo Ambiental (RAMA) en el Valle de México. Estas técnicas han sido usadas desde el siglo pasado; el modelo autorregresivo (AR) y el método para series caóticas. Los valores pronosticados por el modelo caótico son más precisos que el modelo autorregresivo. Se pretende pronosticar los valores máximos de concentración de ozono durante el día. Sin embargo, se requiere estudiar las variantes que tienen estos modelos para tener mayor precisión en el pronóstico, de manera que sea de utilidad práctica para tomar medidas preventivas pertinentes. En esta exposición los algoritmos tienen como principal carencia usar sólo el contaminante ozono en la serie del tiempo, dado que modelos más complejos deberán considerar variables meteorológicas que mejoren la precisión del pronóstico. Así a *grosso modo*, si se considera la serie formada por los máximos diarios de ozono en el período estudiado (enero a mayo de 1999) la desviación estándar es de 0.05 ppm de ozono en la variante más precisa del modelo caótico que se muestra al final del artículo, la desviación estándar del error entre los valores máximos pronosticados y los valores reales es de alrededor de 0.027 ppm de ozono, lo que acorta la banda de valores pronosticados a algo más de la mitad, considerando el total de los máximos diarios como una distribución normal.

**Palabras clave:** Medidas de ozono, Ciudad de México, modelo autorregresivo, modelo caótico.

## Introduction

Atmospheric pollution is one of the most serious problems confronting our modern world [1,2]. In the Metropolitan Area of Mexico City the pollution control is one of its priorities, as it affects the life quality of about 20 million inhabitants as well as its ecosystem.

Air quality in mega metropolitan areas is a function on the quantity and quality of the fuels used, on the technology used by industrial and transportation units, on the high concentration of population and factories, and on the prevalent meteorological condition. The high concentration of emissions and consequently of contaminants is particularly notorious in the metropolitan area of Mexico City. The inhabitants of the metropolitan area are thus critically exposed to those contaminants with a high concentration gradient, as it is the case of carbon monoxide, nitrogen oxides and particles.

The most severe environmental problem in Mexico City is due to ozone and suspended particles. Invisible and with no emission sources, ozone is formed in Nature by complex photochemical reactions involving the ultraviolet spectra of sunlight. The reaction products are ozone and other contaminants.

It is assumed that the level of tropospheric ozone is affected by the meteorological conditions, such as insolation, wind, humidity, temperature, etc. Most of the papers published so far, try to develop a regression model incorporating the meteorological variables to predict an ozone episode. Theory predicts that there is a correlation among the following meteorological variables and the maximum ozone concentration [4]

$$O_3 = f(\text{temperature, humidity, wind speed and direction})$$

Air quality regulations establish the maximum contaminant concentrations to protect the public health and those of more susceptible people, therefore incorporating a safety margin. The regulations have to be observed by local and federal authorities in charge of enforcing environmental programs. In Mexico City, air quality is not satisfactory if the

IMECA (by its Spanish acronym for Índice Metropolitano de la Calidad del Aire) index exceeds 100 at the surface level (equivalent to 0.11 ppm as an hourly average once in a day). The National Air Quality Standard is very often exceeded in a year.

Several programs have been elaborated to control ozone concentration levels in the urban area of Mexico City by the Federal District Government and the Environmental Ministry of the State of Mexico, however, in no one is detailed the meteorological effect on ozone levels. Due to the great amount and variability of conditions that have to do on measuring atmospheric variables and contaminants, it was necessary to filter out the information. Once the information was screened, it was found that data from the Pedregal Environmental Station, has the highest number of high levels of ozone episodes in a yearly basis. Then the surface ozone data during 1999, from Pedregal Station were selected for this study.

One way to study ozone formation in an urban zone is to use the emission inventory of its precursors (nitrogen oxides and volatile organic compounds) from anthropogenic and biogenic sources. In this estimation is important to quantify the intensity of the luminous radiation and to establish the photochemical reactions between the ultraviolet radiation and the ozone precursors. It is also required to determine the intensity and direction of wind fields. Having the above information, it is necessary to solve differential equations to estimate the level of ozone formation, which is quite a complex problem.

Lately, numerous papers have been published advancing equations that relate ozone formation with the meteorological parameters. Results are not satisfactory and the proposed equations vary in accordance to the place of application.

The aim of the present paper is to apply several mathematical techniques to predict the maximum daily ozone level in order to alert the public or to take appropriate measures to reduce its level. Previously, several studies have been made to predict the maximum daily ozone level [2,3,5]. There is not a model universally accepted to predict the maximum ozone level in urban o industrial areas, as the models depend on the climatologically characteristics of the zone considered.

## Methodology

In this first presentation, the surface ozone time series from Pedregal is used. Undoubtedly, other time series from Pedregal should be taken into account to improve forecasting, such as the hourly series of temperature, humidity and other contaminants; such as carbon monoxide, nitrogen oxide and sulfur dioxide. It is possible that the height of the limit layer should be included in the hourly series of barometric pressure; unfortunately, we could not find an electronic register of barometric pressure.

Analyzing the hourly ozone as a time series, the principal methods employed were: the autoregressive (AR) [7], and models based on a simulation of an attractor in an embedded space, this is, considering the series as a chaotic system [6]. In the following paragraphs, the preliminary results on the calculated superficial ozone levels are shown for the urban zone of Mexico City.

## Autoregressive Method

The general form of the predictor is:

$$X(n + 1) = a_0X(n) + a_1X(n - 1) + \ldots + a_qX(n - q) \qquad (1)$$

Where $X(j)$, $j = 1,2,\ldots n$ is a sequence of terms known up to the "$n$" term and the next one is approximated as a linear combination of the "$q$" previous term. The number "$q$" of terms employed to make a prediction is the fundamental parameter on the autoregressive method, and "$q$" is known as the "order" of the method. There is an optimal way of selecting the coefficients "$a_i$" in the equation (1) [7]. A method known as the "Maximum entropy method or All-poles method" was used to predict ozone. A computer program given in reference [7] was used in this paper.

A great advantage of the autoregressive method is its simplicity as it uses only information on the data series on which a prediction is made. On the contrary, the multiple regression methods try to correlate several variables, which for ozone are the series of temperature, humidity, direction and wind speed, giving more complex models.

In order to apply the methods here exposed it is convenient to analyze previously the series, to visualize the trend, periodicity, and in particular to see the mechanism generating the series, in order to select the most appropriate method of prediction.

The selected series is the hourly surface ozone concentration reported in ppm as registered at the Pedregal Environmental Station in the Federal District, in particular the series extending from January to May 1999.

The fundamental aim of this study is to predict with 18 to 24 h in advance the ozone concentration in a meteorological station in Mexico City. As said before, a physical model of superficial ozone formation is a very complex issue; however, the use of a simple model as the AR is a test to visualize its advantage, and also to analyze the statistical consistency of data.

In the first test, data was fed into the AR model as it was registered. In Figure 2, it is shown the ozone levels of 20 consecutive days forecasted with 24 h in advance, from February 23rd to March 14th 1999, which correspond to days 55 to 74 in the data series. The ratio of the mean quadratic error to the quadratic mean value of the signal (hour to hour) is 0.378, equivalent to a correlation coefficient higher than 0.8. In Figure 1, it is illustrated the real maximum ozone values and those predicted during the same 20 days. In this case the ratio of the mean quadratic errors to the quadratic mean value of ozone is 0.327, which approximates to a correlation coefficient of 0.89. In Figure 3 it is shown the hour by hour predicted and real ozone values during the first 10 days.
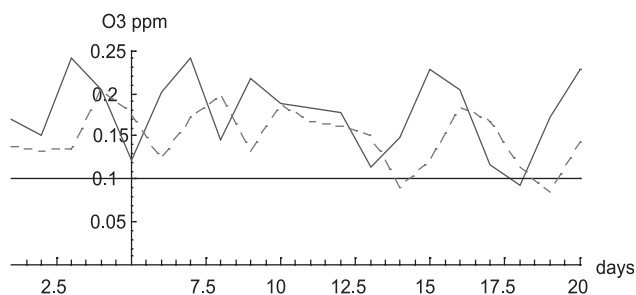
**Fig. 1.** Maximum real and forecasted ozone for a period of 20 days, starting from day 55 of 1999. Real data are depicted by a continuous line and forecasted ones by a dashed line. Forecasting was done by the AR method. Initial set of data comprised 1296 points from first day of 1999, increasing the set each 24 h 60 poles were used.
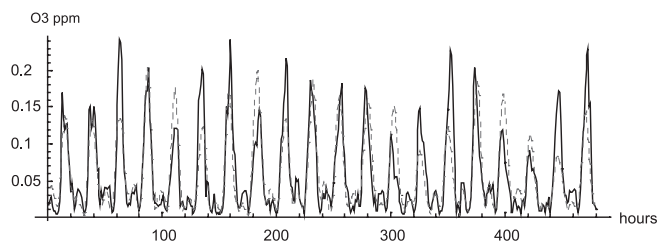


**Fig. 2.** Hourly ozone data. Real values are represented by the continuous line and forecasted ones by the dashed line. The 20 days period correspond to the same one illustrated in Figure 1 and the same method of forecasting was applied.
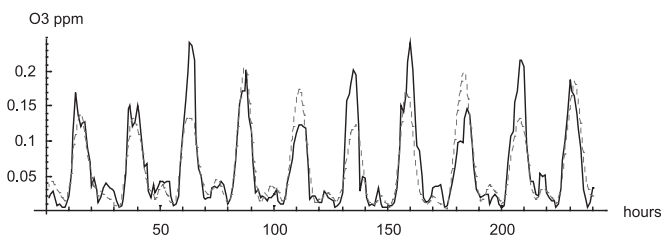


**Fig. 3.** This figure 2 correspond to the first half of Figure 2.

The results are outstanding if it is taken that no previous adjustment to data was made and that the number of poles (number of terms in the linear combination) was not optimized, nor optimized the ratio of poles to number of data to be used in the AR model. Furthermore, forecasting is made with 24 h in advance. It will be shown below that prediction is more precise if it carried out hour by hour.

In order to improve prediction, it was increased the number of poles and the total number of data. Figure 14 shows the result of using 120 poles and around 800 constant points
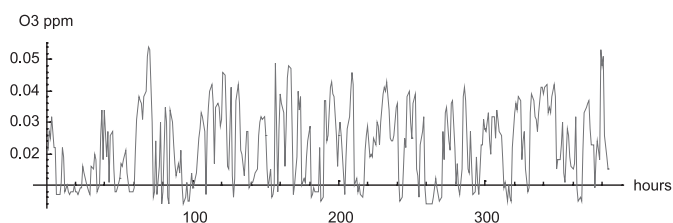


**Fig. 4.** Nightly ozone (from 1 to 6 h) during the first 64 days of 1999. It is appreciated the lack of periodicity and ozone concentration not reaching 0.06 ppm.
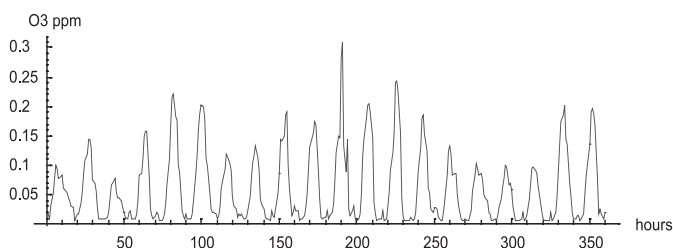


**Fig. 5.** Diurnal ozone at Pedregal Station from 7 to 24 h for the first 20 days of 1999. There is a periodicity and ozone concentration reaches 0.309 ppm.

(796). Prediction improves, as the ratio of the mean quadratic error to the mean quadratic value of the signal for the maximum values forecasted is 0.304. The same ratio for forecasted and real values in an hour by hour basis for the same 20 days was 0.363. However, to improve considerably the forecasting quality, it is necessary to analyze and treat data before applying again the AR method.

Ozone data can be separated in two groups. One is the nightly ozone data from 1 to 6 h, which comprises relatively low ozone values compared with diurnal data. The other group is the diurnal ozone data from 7 to 24 h. In Figure 4, it is shown the nightly ozone concentration data during the first 64 days of 1999. Figure 5 shows the diurnal ozone for the first 20 days of the same period considered. It can be observed that periodicity is complete on the diurnal ozone, as there is a strong correlation between solar radiation and ozone concentration in the atmosphere [4]. This can be appreciated in Figures 6 and 7 that show the correlogram and periodogram of the nightly ozone series during the first 64 days. Correlation is minimum and the periodogram does not show outstanding isolated peaks. On the other hand, the diurnal ozone shows a strong correlation and periodicity as shown in Figures 8 and 9 for the same period of 64 days. For the sake of comparison, in Figure 10 is shown the periodogram side by side of nightly and diurnal ozone. In Figure 11 is shown the total ozone periodogram for the same 64 days.

It is therefore advisable to use the AR method, but with the diurnal ozone data. In Figure 12, it is shown the real maxi-
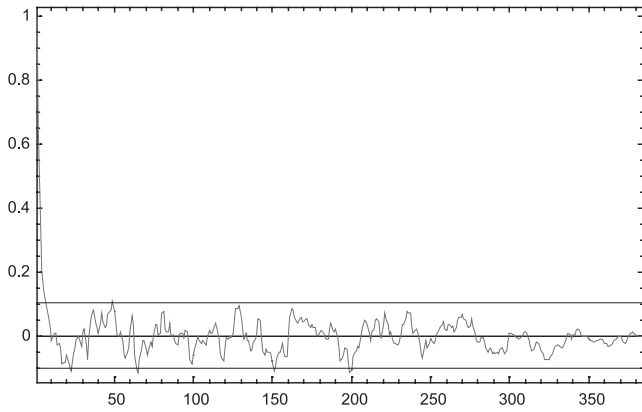
**Fig. 6.** Correlogram of nightly ozone from 1 to 6 h during the first 64 days of 1999 at Pedregal Station.
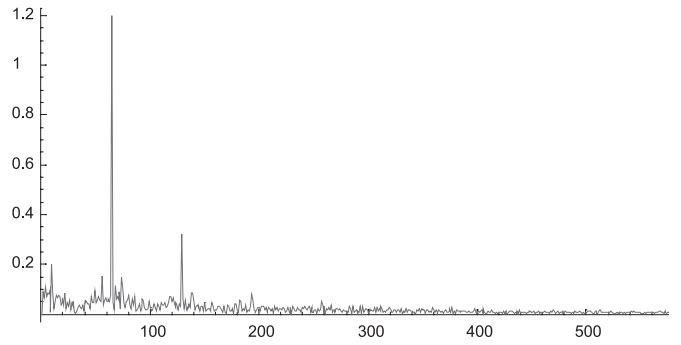


**Fig. 7.** Periodogram of nightly ozone from 1 to 6 h during the first 64 days of 1999 at Pedregal Station.



**Fig. 8.** Correlogram of diurnal ozone from 7 to 24 h for the first 64 days of 1999 at Pedregal Station.



**Fig. 9.** Periodogram of diurnal ozone from 7 to 24 h during the first 64 days of 1999 at Pedregal Station.
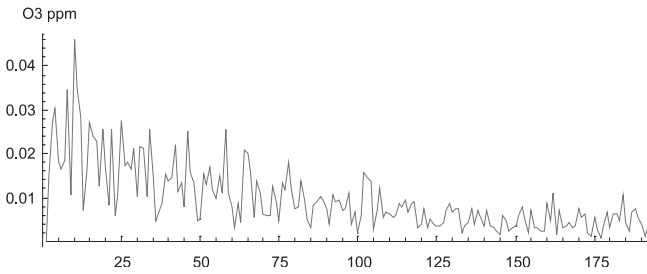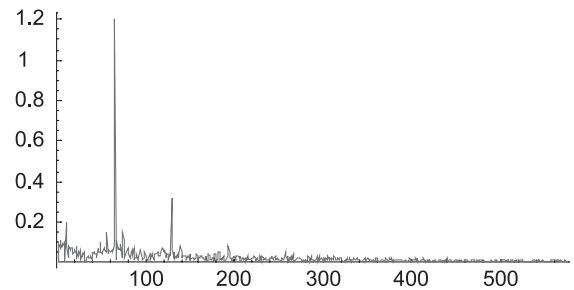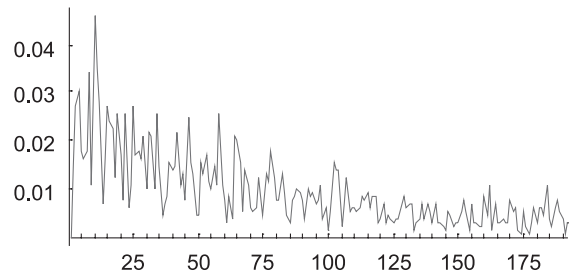




**Fig. 10.** Both periodograms, nightly and diurnal, for the first 64 days of 1999 at Pedregal Station are shown side by side for comparison.
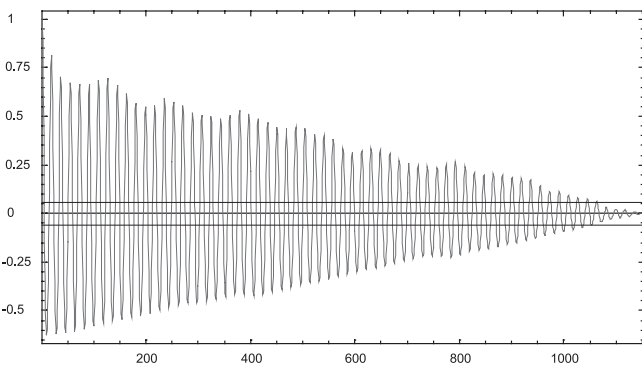
mum and the predicted values for the period of 20 days considered by employing 100 poles, improving the forecast as the ratio of the mean quadratic error to the quadratic mean of signal for the predicted maximum improves, it is now 0.301. The same mean ratio for the hourly prediction is 0.3520.

In other work out with the same AR method, but using 120 poles (in this case is the order of the AR method) the results are shown in Figures 13 and 14. Figure 13 shows simultaneously the real ozone and the forecasted one during the first 10 days, and in Figure 14, it is illustrated the real and forecasted maximum during the 20 days considered before. The ratio of the mean quadratic error to the mean quadratic of signal was 0.3047. Results are shown in Figure 15 for an entirely analogous calculation to the previous one, but considering only 30 poles, giving a ratio of the mean quadratic error to the mean quadratic value of signal of 0.3270.
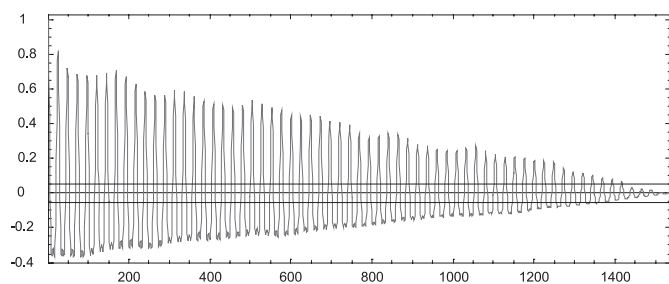
**6** *J. Mex. Chem. Soc.* **2005**, *49(1)*

Margarita Garfias Vázquez, *et al.*



**Fig. 11.** Correlogram of hourly ozone during the first 64 days of 1999 at Pedregal Station.
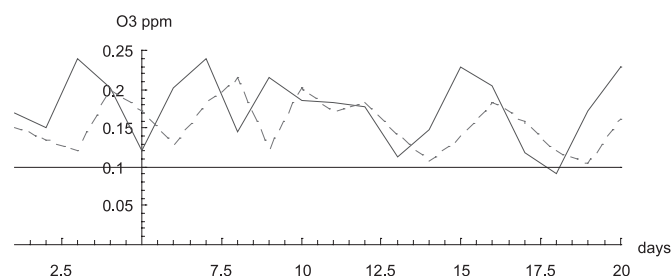


**Fig. 12.** Real and calculated ozone maximum of ozone for a 20 days period, from day 55 of 1999. Real values are depicted by the solid line and predicted ones by the dashed one. The AR method was applied using 100 poles and a constant set of 796 diurnal points from 7 to 24 h, rendering 18 points per day.
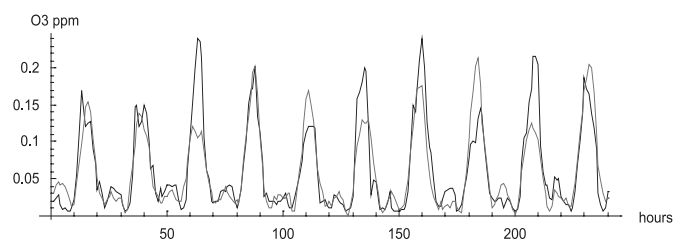


**Fig. 13.** Real and predicted hourly ozone for period of 10 days starting on day 55 of 1999 at Pedregal Station. Real values are shown on a continuous line and predicted ones on a dashed line. The AR method was used to forecast ozone with 120 poles and a constant set of 796 previous points.

## Method of delays

Other method considered for forecasting ozone is the method of delays applied mainly to chaotic models [6]. It is required to generate an embedded space with the ozone series and an affine transformation is used to find the following points in the series. Basically, the time series: $X(ti)$ $(i = 1,2,…n)$ is separated into the following vectors:
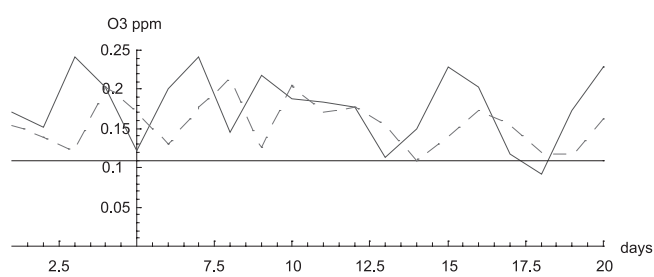


**Fig. 14.** Real and predicted maximum ozone values in a 20 days period from day 55 of 1999. Real values lie on a continuous line and predicted ones on a dashed line. The horizontal line is drawn at a concentration of 0.11 ppm, to show the National Air Quality Standard. The AR method was used to forecast ozone with 120 poles and a constant set of 796 previous points.
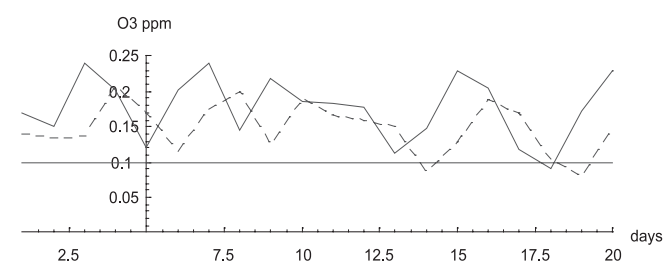


**Fig. 15.** Real and predicted ozone for a 20 days period, from day 55 of 1999. Real maximum ozone values lie on the continuous line and predicted ones on the dashed line. The AR method was used with 30 poles and a constant set of 796 previous points.

$$Xi = \{X(ti), X(ti - T), (X(ti - 2T)… X(ti - (d - 1)T)\} \qquad (A)$$

Where $T$ is the "delay" and "$d$" is the number of coordinates or terms in the vector and the dimension of the embedded space, a vectorial space made precisely by all the vectors in A. To forecast next value, the last of the vectors is taken, for example $x_N$, and then it is looked for in the embedded space for neighboring vectors, which have a succesor $Y_i$ in the series. Be Y the group of all the successive vectors, then a function is built with the vectors $X$ and $Y$. In this case, an affine transformation is adjusted:

$$AX_i + b = Y_i \qquad (B)$$

The set of vectors $X$ and their successors $Y$ are used to adjust by least square method the coefficients of the matrix A and the coordinates of the vector $b$ in equation ($B$). The least square method uses the QR decomposing method or the singular values decomposition method [6]. Once the matrix A and the vector b are found, then the last of the vectors $X_n$ is substituted in ($B$), to obtain its successor one and the first coordinate of this vector is the following predicted datum of the series. The shape of the function $B$ is quite arbitrary, but the one

shown here was selected as it is the must suitable for chaotic models as reported in literature [6,10]. In other circumstances, in a simpler manner, it can be taken the average value of the first coordinate of all the successors in the set Y of vectors [11]. This last technique avoids the need to calculate the A matrix and the vector "*b*", *i.e.* there are no parameters to adjust. It should be noted the great flexibility of the method because of the possibility of selecting at will, a function and the parameters.

Previous to applying the method of delays, the ozone series was smoothed by means of a moving average filter, building a spectral matrix [9]. The matrix was constructed with the first 1200 data in an embedded space of dimension 12 [9]. The moving average data was projected on the first 6 eigenvectors of the spectral matrix [9].

The method of delays was applied on the smoothed values on an hour by hour horizon. Figure 16, shows the predicted maximum values and the real ones in a 20 day period. The
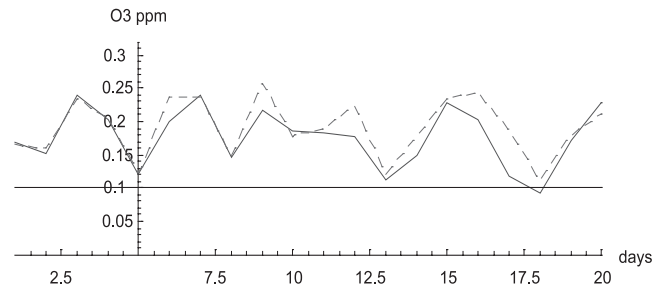


**Fig. 16.** Real and predicted maximum ozone for a 20 days period from day 55 of 1999 at Pedregal Station. Real data are on the continouos line and predicted ones on the dashed line. The method of delays was used with an embedding dimension of 18 and selecting 36 closest neighboring vectors. Data was previosly smoothed by moving averages and it was proyected on the first 6 eigenvectors of the spectral matrix constructed from the first 1200 points, using a window of 12 points. Forecasting was made on an hour by hour basis.
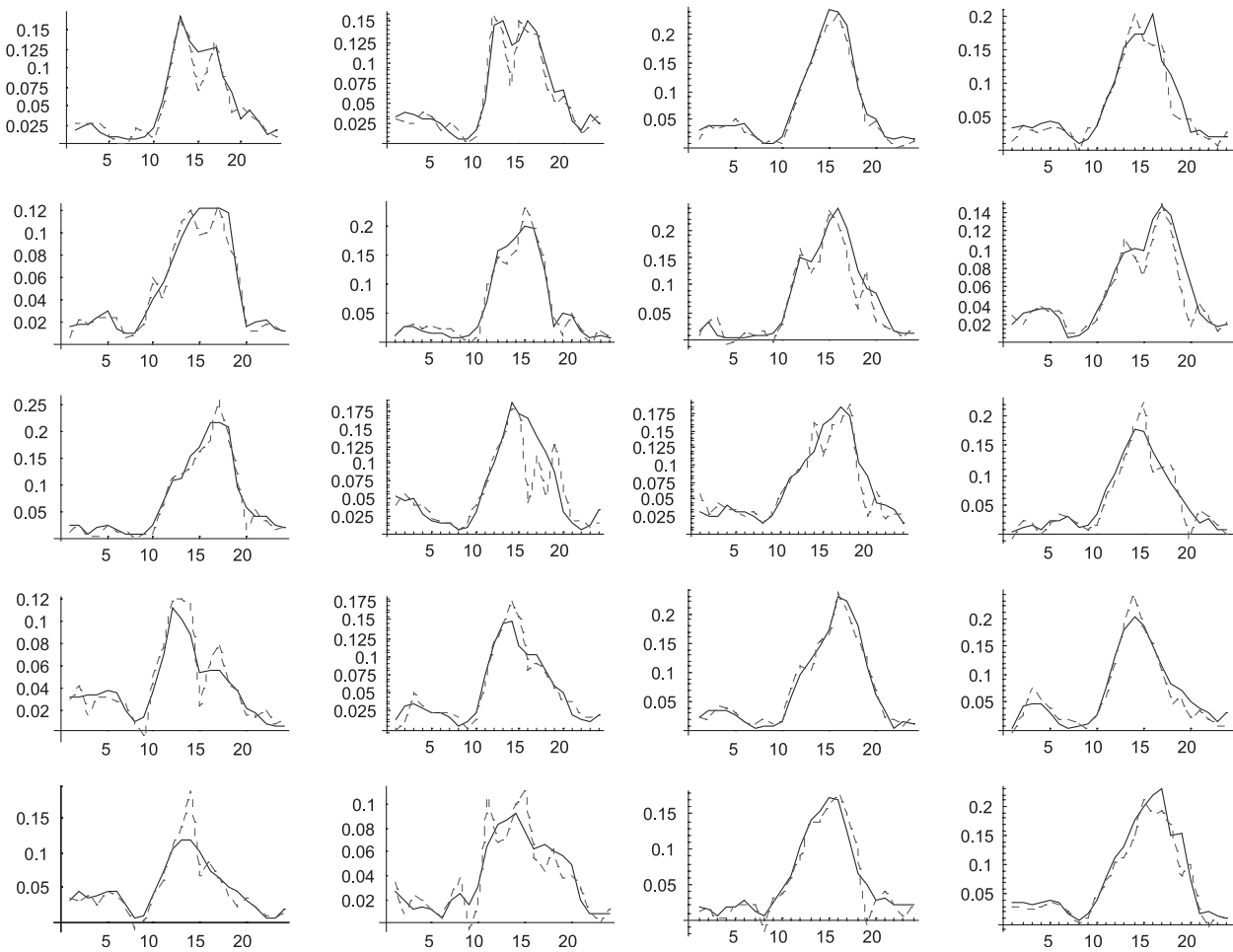


**Fig. 17.** Real and predicted hourly ozone for 20 days from day 55 of 1999 at Pedregal Station. Real hourly ozone on the continuous line and forecasted ones on the dashed line. The method of delays was used to forecast ozone with an embedding dimension of 18 and selecting 36 closest neighboring vectors.  Real data wase previously smoothed as in Figure 16.

forecasted values are quite close to the real ones, as the ratio of the mean square error to the mean quadratic value of the signal is 0.1416, that is equivalent to a correlation coefficient greater than 0.98. In Figure 17, it is shown the notable coincidence of predicted and real values when the forecasting is made hour by hour. The ratio of the mean quadratic error to the mean value of the signal when the prediction is made hour by hour is 0.2233. The low error values are mainly due to the fact that prediction is made by the hour. An embedded space of dimension 18 was used and 36 neighboring vectors were chosen. The results show that the considered ozone series is a chaotic one, notwithstanding that it is not amenable to forecast data well in advance.

To use the method of delays for long prediction, a forecasting is made to an infinite horizon, besides the diurnal ozone was selected (from 7 to 22 h), and smoothed by the moving average filter as before, projecting the series on the first 8 eigenvectors. Prediction starts at 10 h every day. Results are shown in Figure 18 for the maximum values in the same 20 days, and the forecast is outstanding as the ratio of its mean quadratic error to the mean quadratic value of the signal

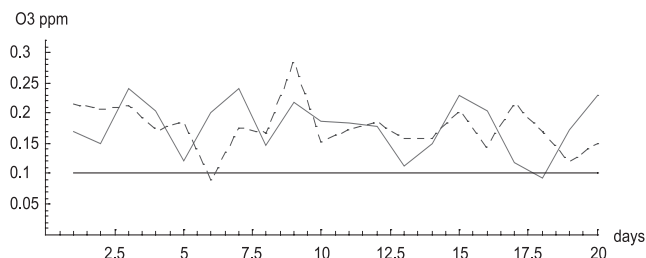is 0.1634. In Figure 19 is shown the prediction hour by hour in the first 12 days.



**Fig. 18.** Real and predicted maximum ozone for a 20 day period (Monday to Friday) for the 4 weeks from week 17 of 1999, without considering half a week at the beginning of the year.The delay method was applied with an embedding dimension of 16. Two closest neighboring vectors were selected and arithmetic averaging was used to forecast next values. Data were previously smoothed with a dynamical average [9] on the first 8 eigenvectors of a spectral matrix constructed on the same embedding space.
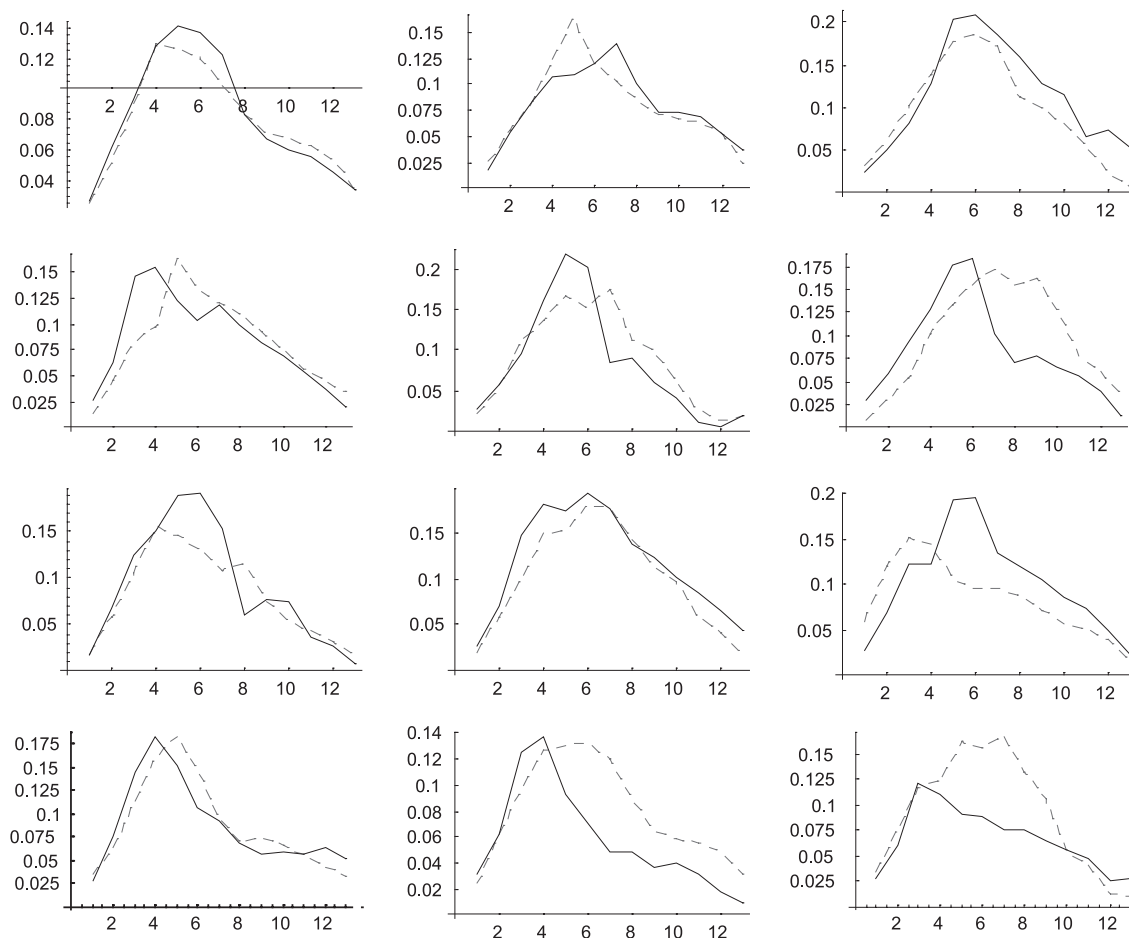


**Fig. 19.** Graphical array showing the hourly ozone from 10 AM to 11 PM, Real and predicted hourly diurnal ozone for the first 12 days of the 20 weekly days reported in Figure 18.Real ozone lies on the continuous line and forecasted one on the dashed line.

## Conclusions

It is here shown that those elementary methods such as the AR and the chaotic one, in which the only information they contain is the series itself, are capable of giving satisfactory forecasting on the surface ozone concentration. The trends on the predicted values are close to the real ones. In the AR method, increasing the poles number used improves precision. It is clear that the hourly ozone levels depend strongly on the previous values. An outcome of this study is validation of data measured at Pedregal Environmental Station. In a second part, meteorological variables will be introduced to improve ozone forecasting.

## References

1. Garfias, F.J.; Díaz, L. *Gasolinas Oxigenadas: La Experiencia mexicana*, 1ª. Edición, **2003**, Fondo de Cultura Económica, México.
2. INE, Almanaque de datos y tendencias de la calidad del aire en ciudades mexicanas, **2004**, México.
3. Pao-Wen Grace Liu, *Forecasting Peak Daily Ozone Levels I. A Regression with Time Series Errors Model Having a Principal Component Trigger to fit 1991 Ozone Levels*, J. Air & Waste Manag. Assoc., **2002**, *52*, 1064-1074.
4. Brown, M. J. Mexico City Ozone Concentration as a Function of Readily Available Parameters, February **1994**, Los Alamos National Laboratory, Report 87545, Los Alamos,
5. Comrie, A. C., *Comparing Neural Networks and Regression Models for Ozone Forecasting*, *J. Air & Waste Manag. Assoc.*, **1997**, *47*, 653.
6. Mattheij, R.M.M.; Molenaar, J., *Ordinary Differential Equations in Theory and Practice*, **1996**, J. Wiley and Sons, New York.
7. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T., "Numerical Recipes: The Art of Scientific Computing (FORTRAN Version)" **1989**, Cambridge University Press, Cambridge.
8. Wolfram Research Inc., "Mathematica", **2004**.
9. Shaw, W.T.; Tigg, J. *Applied Mathematica: Getting Started, Getting it done*, **1994**, Addison Wesley Pub. Co., New York.
10. Farmer, J.D.; Doyne, J.; Sidorowich, J.J., *Phys. Rev. Letl*. **1987**, *59*, 845.
11. Jian-Long Ch.; Islam, S.; Biswas, P. *Atm. Envir.* **1998**, *32*, 1839-1848.