Estimation of daily surface ozone using periodic and stochastic modeling in Chennai region

K Padma$^5$, R Samuel Selvaraj & B Milton Boaz
Presidency College, Chennai, 600 005, India
$^5$E-mail: padma_mamalai@rediffmail.com

Received 2 June 2014; revised 22 September 2014; accepted 1 October 2014

The present study deals with the modeling and forecasting of surface ozone time series in an urban area. First, an analysis of the systematic components (periodicity and stochastic components) was performed. Subsequently, prediction model for the daily surface ozone series was developed. In the recent years, there was no permanent measurement of surface ozone at this site, so surface ozone was measured from June 2011 to September 2012 at the urban site Koyembedu, Chennai (the capital of Tamil Nadu), India. Daily cumulative ozone data series was obtained by hourly instantaneous data. It was found, using Mann-Kendall test, that the data series is free of trend. The periodicity of ozone data was analyzed using Fourier Transform method. Stochastic components of ozone data are assumed as residues between observed ozone data and values computed from periodic model. Stochastic model presented in this research is basically a 3rd order autoregressive model. The developed models were validated using correlation coefficient between the predicted values and measured values. The spectrums of series exhibit 100 days period of daily surface ozone and implies that the pattern in the series is repeated every 100 days. The correlation coefficient (R) of this model delivers 0.810 and can provide mean bias error (MBE) = 0.85, and root mean square error (RMSE)=0.83. The result suggests that this approach is good for estimating daily surface ozone with sufficient accuracy.

Keywords: Surface ozone, Periodic modeling, Stochastic modeling, Fourier transform, Autoregressive model

PACS No.: 92.60.Sz

1 Introduction

Air pollutants have environmental impact on public health, properties, vegetation, etc. To prevent or minimize the damage caused by atmospheric air pollution, suitable monitoring systems are urgently needed to measure and quantify pollutants concentration by regulating authorities in order to prevent further deterioration of the current air quality status. Air pollution is a serious problem in thickly populated and industrialized urban area. The air pollution is high, especially in areas where pollution sources and the human population are concentrated. Ozone is a secondary pollutant resulting from photochemical reaction of a variety of natural and anthropogenic precursors mainly volatile organic compounds (VOCs) and oxides of nitrogen (NOx = NO + NO2). Under favourable meteorological conditions, ozone may accumulate in the atmosphere and reach such a high concentration level that can impose adverse effects on human health and ecosystem. The concentration of ozone in the troposphere is of great interest because of its negative influence on human health, vegetation and materials. Accumulation of ozone near the ground is influenced by physical and chemical processes and by meteorological conditions. Both industrial and highly populated areas are the main sources of ozone precursors. High concentration of ozone is recorded usually on the outskirts of urban areas or a few kilometers from industrial areas. This is due to the photochemical reaction taking place during the transport of the precursor. Development and use of statistical and other quantitative methods in the environmental sciences have been a major communication between environmental scientists and statistician. In recent years, many statistical analyses used to study air pollution as a common problem in urban areas. Surface ozone (O3) is the most important air pollutant causing damage to vegetation including crop yield and forest productivity, losses and alterations to net primary productivity. In order to understand and control pollutant level, some models need to be developed.
The time series consists of a set of sequential numeric data taken at equal intervals, usually over a period of time or space. The first step in time series analysis is to draw time series plot which provides a preliminary understanding of time behaviour of the series. Time series analysis is a useful tool for better understanding of cause, effect and relationship of environmental pollution. The main aim of time series analysis is to describe movement history of a particular variable in time. Many researchers have tried to detect changing behaviour of air pollution through time using different techniques.

Management of control and public warning strategies for pollutant levels at densely populated areas requires accurate forecasts of ambient level, although ozone prediction models have been proposed at several cities. One should know the pollutant surface ozone in conjunction with time series and it is very difficult to prove a series of record keeping of ozone. Accumulation of surface ozone is influenced by physical and chemical processes and by meteorological conditions. So, in order to ascertain ground level ozone, one should measure all the influencing factors along with the ozone. Only then, one can understand the complex variability of ozone over a wide range of spatial and temporal scale. This underlying factor makes it very difficult to keep the measure of surface ozone on a large scale because it requires a lot of manpower and also involves huge expenditure. So, sometimes to predict or add the data recording of ozone, synthetic ozone data is required. Various methods have been developed by researchers in the field of engineering and science to prove this information. Most widely used methods are deterministic and stochastic methods.

A time series can be described either in the time domain or in the frequency domain. Any time-varying signal can be constructed by adding together sine waves of appropriate frequency, amplitude and phase. Fourier Transform is a technique that is used to determine which sine waves a given signal is made of, i.e., to decompose the signal into its constituent sine waves. The result is expressed as sine wave amplitude as a function of frequency. Thus, in the power spectrum, a signal consisting of a single sine wave will have a peak matching the frequency of the sine wave, and zero contributions from all other frequencies.

In order to reduce the number of computations, different algorithms were developed by various researchers. In the present study, power spectral analysis is done using Fourier Transform method, which is implemented using Fast Fourier Transform algorithm. The most widely used algorithm for computing the discrete Fourier transform (DFT) is the Fast Fourier Transform (FFT). The number of computations involved in this procedure is much lesser than that in the direct computation of DFT. The computations in direct estimation of DFT are of the order of \( N^2 \), which is reduced to the order of \( N \) (\( \log N \)) if FFT algorithm is implemented. This algorithm has been implemented using MATLAB software. The size of the data array, \( N \) must be a power of two to achieve faster computation, otherwise the computational speed will be reduced. The execution time of an FFT algorithm depends on the transform length. It is fastest when the transform length is a power of two, and almost as fast when the transform length has only small prime factors. It is typically slower for transform lengths that are prime or have large prime factors. Time differences, however, are reduced to insignificance by modern FFT algorithms such as those used in MATLAB. Adjusting the transform length for efficiency is usually unnecessary in practice. For an array of real numbers of size \( N \), FFT implementation returns an array of equal size, \( N \) points, but only the first \( N/2 \) are valid values. The second half will be the mirror image of the first half. One can compute the power as a function of frequency \( \nu \).

Due to the periodic motion of the earth around the sun and about its own axis, rainfall like any event is believed to have periodic components. Moreover, some researcher suggested that the monthly rainfall sequence is considered to be a periodic-stochastic process. The ozone has periodic and stochastic parts in nature, because ozone is influenced by weather parameters. From the surface ozone, time series can determine both periodic and stochastic components. Hence, analysis of periodic and stochastic ozone time series will produce a model that can be used to calculate the periodic and stochastic data and can also be used to predict the daily surface ozone variation.

In the present study, an attempt has made to develop periodic and stochastic models that estimate the surface ozone concentrations in air in an urban area in south India.
2 Study area

The study is carried out at Chennai, situated on the south east coast of India and north east coast of Tamil Nadu, is one of the highly populated urban site. Chennai lies on the thermal equator and is also a coastal city (Fig. 1). The latitude and longitude of the center of the city are 80°14'51"E and 13°03'40"N. Chennai can be divided into four areas: North, Central, South and West. The northern part is primarily an industrial area comprising of petrochemical industries in the Manali area and other general industries in Ambattur. Southern Chennai also has an industrial area (Guindy Industrial Area) and the Western part near Sriperumbudur on Bangalore Highway is being developed with new industries. Manali area, northern part of the city, is the home to petroleum refinery (Chennai Petrochemical Corporation Limited or CPCL) and allied petrochemical industries. Ambattur area houses several small scale industries while the western area of Sriperumudur houses different automobile industries as Hyundai, Ford, etc. Surface ozone was measured throughout Tamil Nadu during the year 2011 and it was found that Kanniyakumari district had the highest daily average of 17.8 ppbv. Moreover, it was found that the hourly values of the surface ozone levels, studied in Chennai during 2004-2005, varied from 1 to 50.27 ppbv. In the present study, the surface ozone (O$_3$) concentration was measured in the urban site, Chennai. This study was conducted at Koyembedu, Chennai, which houses Chennai’s Moffussil Bus terminus and hundreds of buses and other vehicles ply daily and hence, the vehicular emission is very high. This site is surrounded by a number of industrial areas located within a short radius. Further, a waste water purifying plant is also located in the close proximity.

3 Data collection and Method

Hourly instantaneous surface ozone and all other corresponding meteorological parameters had been measured from June 2011 to September 2012 and used for analysis. Night time measurement had been made on some days of the month and it was found that surface ozone levels were very low during the nights (<10 ppbv). The hourly values and daily average values were used to analyze the day-to-day variability. Surface ozone measurements consist of 446 daily observations, daily ten measurements carried out on all days between 08.00 and 17.00 hrs IST. An Aeroqual series 200 ozone monitor was used to measure low and high surface ozone levels.
Its low concentration ozone head measures the ozone concentration from 0.000 to 0.500 ppm and a high concentration ozone head measures the ozone concentration from 0.050 to 20.00 ppm. Accuracy of a low concentration ozone head is ±0.010 ppm (from 0 to 0.100 ppm); ±10% (0.10 to 0.500 ppm). While that of a high concentration ozone head is ±10% (from 0.20 to 2.00 ppm); ±15% (from 2.00 to 20.00 ppm), the measurement units being either ppm or mg m\(^{-3}\). The sensor type is gas sensitive semiconductor (GGS) and works on the principle of absorption of UV radiation by ozone in the ambient air. The ozone sensor was calibrated against a certified UV photometer. The particular instrument was chosen for its simplicity and reliability in operation; ease of handling; and cost effectiveness and speed in obtaining gas concentration level directly. This has been a great assistance in estimating the concentration of ground-level ozone in places where there was no permanent measurement. An Aeroqual monitor with GSS ozone sensors were used by several researchers for the measurement of the atmospheric ozone and nitrogen dioxide\(^{19-21}\). The principal aim of the analysis was to obtain a reasonable model for estimating original data series into its various components. Generally, a time series can be decomposed into a deterministic component, which could be formulated in a manner that allowed exact prediction of its value, and a stochastic component, which is always present in the data and cannot strictly be accounted for as it is made by random effects. The time series, \(X(t)\), is represented by a decomposition model of the additive type\(^{22,23}\) as:

\[
X(t) = T(t) + P(t) + S(t) 
\]  

(1)

where, \(T(t)\) is the component of trend \(t = 1, 2, 3, \ldots N\); \(P(t)\), the periodic components; and \(S(t)\), the stochastic components.

### 3.1 Mann-Kendall Test

There are several approaches for detecting the trend in the time series. These approaches can be either parametric or non-parametric. Parametric methods assumed the data should be normally distributed and free from outliers. On the other hand, non-parametric methods are free from such assumptions. The most popularly used non-parametric tests for detecting trend in the time series is the Mann-Kendall (MK) test. It is widely used for different climatic variables.

The Mann-Kendal (MK) test searches for a trend in a time series without specifying whether the trend is linear or nonlinear. It is based on the test statistics \(S\), which is defined as:

\[
S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(x_j - x_i) 
\]  

(2)

where, \(x_i\), are the sequential data values; \(n\), the length of the data set and

\[
\text{sgn}(t) = \begin{cases} 
1, & \text{for } t > 0 \\
0, & \text{for } t = 0 \\
-1, & \text{for } t < 0 
\end{cases} 
\]  

(3)

The value of \(S\) indicates the direction of trend. A negative and positive value indicate falling and rising trend, respectively. Mann-Kendall have documented that when \(n\) is greater or equal to 8, the test statistics \(S\) is approximately normally distributed with mean and variance as follows:

\[
E(S) = 0 
\]  

(4)

\[
\text{Var}(S) = \frac{1}{18} \left[ n(n-1)(2n+5) - \sum_{i=1}^{n} f_i (t_i -1)(2t_i + 5) \right] 
\]  

(5)

where, \(m\), is the number of tied groups; and \(t_i\), the size of the \(i\)th \(t_i\), group. The standardized test statistics \(Z\) is computed as:

\[
Z_{mk} = \begin{cases} 
\frac{s-1}{\sqrt{\text{Var}(S)}}, & \text{for } S > 0 \\
0, & \text{for } S = 0 \\
\frac{s+1}{\sqrt{\text{Var}(S)}}, & \text{for } S < 0 
\end{cases} 
\]  

(6)

\(P\) is the hypothesis test of slope parameter in MK test, which tells that a statistic called \(P\) that is linearly related to a non-parametric test \(U\) and widely used in studies of pattern recognition. The use of statistical tests involves testing of the null hypothesis, which assumes that the data are random and are not correlated. Another alternative hypothesis is a significant trend detected with either positive or negative. If the level of significance is 0.05, then a p-value > 0.05 should accept the null hypothesis (no trend). In this study, XLSTAT software is used to detect the trend. The standardized Mann-Kendall statistics \((Z)\) follows the standard normal distribution with zero mean and unit variance. Here, the null hypothesis about no trend is accepted as
$P$ value (0.47) is greater than the significance level (> 0.05). Hence, through this study, the ozone data is assumed to have no trend. So, Eq. (1) can be presented as:

$$X(t) = P(t) + S(t) \quad \ldots (7)$$

This Eq. (7) is for obtaining the periodic and stochastic models of daily cumulative ozone series.

### 3.2 Spectral method

Spectral method is one of the several transformation methods and is used in many applications. This method can be presented as Fourier transformation method\(^9,21,22\) as:

$$P(f_m) = \frac{\Delta t}{2\pi} \sum_{n=-N}^{N} P(t_n) e^{-2\pi i(m,n)} \quad \ldots (8)$$

where, $P(t_n)$ is a series of surface ozone in the time domain and $P(f_m)$ is a series of surface ozone in the frequency domain. On the basis of surface ozone frequency resulted from Eq. (8), power is a function of the ozone frequency and can be generated. The frequencies of surface ozone from significant power (amplitude\(^2\)) used to simulate synthetic surface ozone are assumed as dominant surface ozone frequencies. The dominant frequencies resulting in this study, in the form of angular frequencies, were used to determine periodic surface ozone components.

### 3.3 Periodic component

The existence of $P(t)$ is identified by using the Fourier transformation method. This part that shows the presence of an oscillating $P(t)$, using a period $P$ (period = 1/frequency), some frequencies with peak amplitudes can be estimated by using Fourier transformation. Then, an equation can be expressed in the form of the Fourier series\(^1\)as:

$$\hat{P}(t) = S_o + \sum_{r=1}^{k} A_r \sin(\omega_r t) + \sum_{r=1}^{k} B_r \cos(\omega_r t) \quad \ldots (9)$$

where, $P(t)$, is the periodic component; and $\hat{P}(t)$, the periodic component of the model. By using the least squares method, $A_r$ and $B_r$ coefficients of Fourier components can be obtained.

### 3.4 Stochastic components

The periodic component $P(t)$ concerns an oscillating movement, which is respective over a fixed interval of time. By using the results of the simulations obtained from periodic ozone models, stochastic components $S(t)$ can be generated. The stochastic components are the difference between observed ozone data series and values computed from periodic model. Stochastic series as a residual ozone series, which can be presented as:

$$S(t) = X(t) - P(t) \quad \ldots (10)$$

Stochastic components are formed by a random value that cannot be calculated precisely. In the case of ozone time series, various meteorological parameters and precursors response to the value of this component without changing the cycle itself and thus add randomness to the time series. This model was applied to the $S(t)$, which was treated as a random variable, i.e. deterministic components were removed and the residual was stationary in nature. So, stochastic models, in the form of autoregressive model can be written as the following mathematical functions\(^16\):

$$S(t) = \varepsilon + \sum_{r=1}^{P} b_r \cdot S(t - r) \quad \ldots (11)$$

The Eq. (11) can be decomposed into:

$$S(t) = \varepsilon + b_1 \cdot S(t - 1) + b_2 \cdot S(t - 2) + \ldots + b_P \cdot S(t - P) \quad \ldots (12)$$

where, $b_r$, is the parameter of the autoregressive model; $\varepsilon$, the constant of random numbers; $r = 1,2,3,4\ldots$; and $P$, the order of stochastic components. To get the parameter of the autoregressive model and the constant of random number, regressive tool in MATLAB is used.

### 4 Results and Discussion

Surface ozone measured during June 2011-September 2012 at the urban site Chennai had been taken to study the daily time series. Figure 2 shows the daily surface ozone time series of Chennai location. It is a surface ozone concentration with respect to daily time scale. It is clear that day average of ozone concentration varied gradually from 7 to 30 ppbv and the highest peak 30 ppbv was attained in the summer season. This may be due to the local meteorological factors and the activities involved in urban areas where vehicles are the major contributors and more industrial areas are surrounded over the study site.

On the basis of daily surface ozone data and by the method of Fast Fourier Transform (FFT), a power spectrum of daily ozone time series is generated.
To estimate the presence of the periodic component of the daily surface ozone time series, Fourier transformation method was applied to get dominant frequencies of the ozone. To produce the dominant frequencies, FFT algorithm is used by MATLAB, where the amount of the data $N$ is analyzed as the power of 2, for example $N = 2^k$.

In the present work, 446 data series is used, out of which 400 data series is used for calculating periodic and stochastic components and 46 data series is used for testing the result. The transformation result of the daily surface ozone data over a period of 400 days was calculated by the method of FFT (Fig. 3).
Figure 3 illustrates that the spectrum of series exhibits first dominant highest peaks at 0.01, 0.025 cycle per day. The sharp peak indicates a significant amount of the variance having a 100-days periodicity. A smaller peak with a 40-days periodicity is also observed in the ozone series. This shows that the 100-days period of daily ozone data are very dominant in comparison to other periods. This implies that the pattern in the surface ozone data is repeated every 100 days. This is obtained using frequency on the x-axis and power density in y-axis. From the spectrum, it is inferred that the noted major peaks and its corresponding x-axis value gives frequency and inverse of this value gives period. The spectrum presented in Fig. 3 is produced using the FFT method of MATLAB version 2008. This frequencies value substituted in the general equation.

\[ P(t) = a+b \sin(f_1 t) + c \sin(f_2 t) + d \cos(f_1 t) + e \cos(f_2 t) \]  
(13)

As for this problem, the first value \( f_1 \) is 0.01, and second value \( f_2 \) is 0.025, and ‘t’ value is taken as observed ozone value \( t_1, t_2, \ldots, t_n \). The unknown coefficient values \( (a, b, c, d, e) \) are found by MATLAB curve fitting tool (cftool). In cftool, ‘X’ is taken as 1 to 400 real number and ‘y’ is taken as 1 to 400 ozone data. Using the Eq. (13) and parameter values, periodic component was found. Using the results obtained from the generalized Eq. (13), periodic component of this model is found. Residual of daily surface ozone is obtained by the difference between observed ozone series and the periodic ozone series. The residual series of ozone are errors of the periodic ozone model. The periodic errors are assumed as stochastic components of the daily ozone. The stochastic components of the daily ozone are shown in Fig. 4(b).

Autocorrelation method is used to find the order of the stochastic component of the series. That is, the first least value of autocorrelation which (near zero) gives order of the series. This is shown in Fig. 4(a) and stochastic component of autocorrelation series is shown in Fig. 4(b). Moreover, time lag of -4 is found in the present study and the order of the series is estimated to be 3. Using the order the stochastic components, model of daily surface ozone equation is formed.

\[ S(4) = a_0 + a_1 S_1 + a_2 S_2 + a_3 S_3 \]  
(14)

\[ S(5) = a_0 + a_1 S_2 + a_2 S_3 + a_3 S_4 \]  
(14a)

\[ S(6) = a_0 + a_1 S_3 + a_2 S_4 + a_3 S_5 \]  
(14b)

The regression tool in MATLAB (“regress(X,Y)”) has been used. The found model parameters \( (a_0, a_1, a_2, a_3) \) are shown in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0</td>
<td>-0.001</td>
</tr>
<tr>
<td>a1</td>
<td>0.02423</td>
</tr>
<tr>
<td>a2</td>
<td>0.22964</td>
</tr>
<tr>
<td>a3</td>
<td>0.25832</td>
</tr>
</tbody>
</table>

Table 1 — Parameters of 3rd order auto regressive model

Fig. 4 — (a) Residual autocorrelation function (ACF) plot; (b) Stochastic components of daily surface ozone series
Using the parameters in Table 1, the stochastic component 401 to 446 of daily surface ozone can be simulated using Eq. (14) as:
\[ S_{401} = a_0 + a_1 S_{398} + a_2 S_{399} + a_3 S_{400} \]  
\[ S_{402} = a_0 + a_1 S_{399} + a_2 S_{400} + a_3 S_{401} \]  
\[ S_{446} = a_0 + a_1 S_{443} + a_2 S_{444} + a_3 S_{445} \]  

Using the parameters of autoregressive model and the independent random number in the Table 1, the stochastic component of daily surface ozone can be simulated and is presented in Fig. 5(a). Also, simulated periodic component and measured surface ozone are shown in Fig. 5(b).

Figure 6 shows that comparison between the measured and the calculated monthly daily surface ozone of the periodic and stochastic modeling. The figure indicates that the calculated daily surface ozone of the periodic and stochastic model gives highly accurate result. The modeling of the periodic surface ozone alone provides the value of correlation coefficient R as 0.723. Moreover, the stochastic modeling of surface ozone using 3rd order autoregressive model alone gives the value of correlation coefficient R as 0.745. But the result of stochastic and periodic surface ozone model when studied together gives the value of correlation coefficient between the measured data and the model as 0.81. This shows that the periodic and stochastic ozone model when used together has given almost close to the value of measured daily surface ozone data.

Finally, on checking the above presented results, it is found that the 3rd order autoregressive model...
can give better accuracy of results than the 4th order accuracy result. In the present study, stochastic model is created using the 3rd order accuracy. The correlation coefficient (R) and error (%) for periodic and stochastic model delivers 0.810, 6.38% and can provide error values MBE = 0.85, RMSE = 0.83. In the present study, with 400 periodic components, modeling of daily surface ozone can be done quickly. It is because with the method of Fast Fourier Transform (FFT), prediction of the frequency components of harmonic daily ozone can be generated quickly. The result of periodic and stochastic models (P + S) of daily surface ozone when studied together has given a very good correlation and accuracy.

5 Conclusions

In this paper, using measured daily surface ozone series, fast Fourier transform, autoregressive model, the method of least squares are applied, which can produce synthetic ozone series significantly. The spectrum of the daily ozone time series generated by the FFT method is used to simulate the synthetic daily ozone series with less error. In the future work, using the same method, other ozone series can be used to study periodic and stochastic ozone components, and other spectrum methods can be used to generate more accurate ozone spectrums. The performance of this model suggests that this kind of modeling can be suitable for ozone concentrations forecasting.

Acknowledgement

The authors would like to thank the Tamil Nadu Pollution Control Board, Chennai for providing valuable guidance and for checking the accuracy of the instruments. They also thank their co-researchers who have given their valuable suggestions and their help is gratefully acknowledged.

References

6. Emberson Lisa, UK ground level ozone: A threat to food security and a contributor to multiple stresses experienced by plants (Environment Institute, University of York, Harlington, York), 2000, 10.