Box Jenkins methodology applied to the environmental monitoring data

Mihaela Mihai and Irina Meghea

Abstract. In time series analysis, the Box-Jenkins methodology applies autoregressive moving average ARMA models to find the best fit of a time series to past values of this time series, in order to make forecasts. This paper applies the Box-Jenkins methodology to modeling and analysis of the CO monitoring data measured by A.P.M. Bucharest in some important crossroads of Bucharest during 2005 - 2009. We determine the ARMA model that correlates the data without trend and seasonality.

M.S.C. 2010: environmental monitoring, Box-Jenkins methodology, air pollution, time series.
Key words: 62P12, 91B82, 62M10, 60G17, 60G25, 62M20.

1 Introduction

Various aspects referring to statistics of environmental monitoring data applied in evaluation of air pollution are developed by the authors in a series of papers [5]-[9] and a monograph [4]. We can see other statistical methods in [11] and [12].

In time series analysis, the Box-Jenkins methodology applies autoregressive moving average ARMA or ARIMA models to find the best fit of a stationary time series, in order to make forecasts. A time series, TS, is a sequence of observations which are ordered in time. If observations are made on pollutant concentration throughout time, the data are displayed in the order they arose, since successive observations probably are dependent. The environmental air monitoring data obtained in several main intersections from Bucharest form discrete TS. Such TS consists in several tens thousands of registered data, and their graphical representation is actually impossible to be performed. The information presented by diagrams is particularly increased in a 3D representation. As an example, in the figure 1 is illustrated the 3D diagram of CO concentrations in the air in Bucharest intersection Mihai Bravu; the data are collected by A. P. M. - Bucharest. The use of two time coordinates, hours and days, is highly suggestive in underlying the qualitative features of TS. In the intersection analyzed, the circadian rhythm of road traffic is mainly represented by the cyclic fluctuations of CO concentrations in the air as reveals the response surface from Fig. 1.

Applied Sciences, Vol.13, 2011, pp. 74-81.

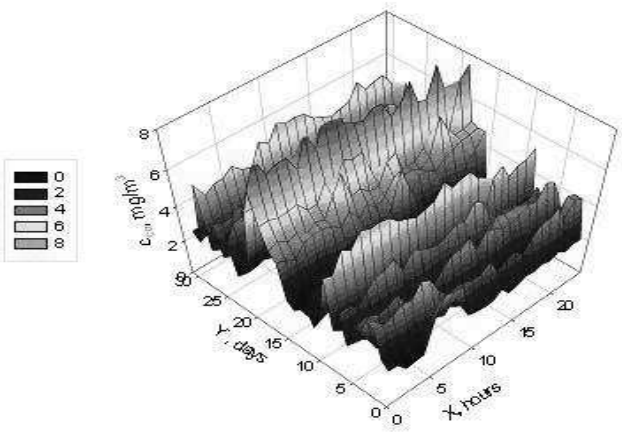

Fig. 1. CO concentration in the air during March 2007 in Bucharest, Mihai Bravu intersection; color code shows the CO concentration in mg/m^3.

2 Application of Box - Jenkins methodology

2.1 First step of Box - Jenkins methodology

The first step in Box - Jenkins methodology is to prove that the collected data form stationary TS with the following properties: homogeneity, variability, periodicity, and interdependence [2, 3, 10].

TS of pollutant concentrations are homogeneous because its terms are similar in their nature and are measured in the same conditions.

The variability of TS terms arises from the tendency that each term is obtained by measuring individual data with corresponding changes determined by the random factors and the dynamics of pollution sources. In the case analyzed, the main pollution source is given by the exhaust gases from road vehicles. As a result, the CO concentration in the air will be dependent on the road traffic intensity, technical state of vehicles, quality of the commercial fuels, etc.

The periodicity or cyclic character of the analyzed TS was put in evidence by the response surface from figure 1. This diagram clearly shows the daily and the weekly periodicity of TS.

The interdependence of TS terms results from the succession of measurements. The Hurst method put in evidence the interdependence of the data from TS [16]. For different data sets, the value of Hurst coefficient for the TS analyzed is ranging between 0.6 and 0.8. This shows that the value of CO concentration in air measured at a given moment is influenced by the value measured before, and therefore the data form TS.

Stationary TS is one whose its statistical properties such as mean, variance, and autocorrelation are all constant over time. As a result, stationary TS have no trend and periodicity. The data analyzed in figure 2 shows that during 2007, the TS is stationary, because the coefficient of linear variation has negligible value as shows the following equation obtained by linear regression of the data: $c = 1.89 + 9.2\ 10^{-6}t$.

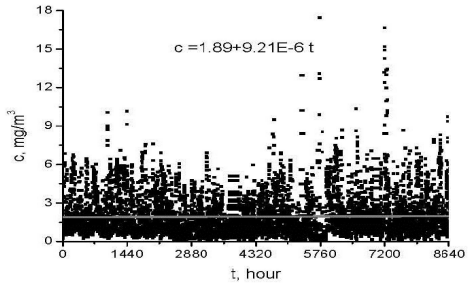

Fig. 2. CO concentration in the air during 2007 in Bucharest, Mihai Bravu intersection.

The autocorrelation diagram from Fig. 3 puts in evidence the periodicity of the data and as result the CO TS data set is non stationary.

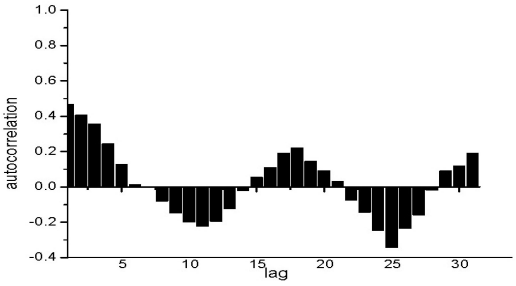

Fig. 3. Autocorrelation diagram of CO concentration in the air during May 2007 in Bucharest, Mihai Bravu intersection.

Most statistical forecasting methods are based on the assumption that the TS can be rendered approximately stationary through the use of mathematical transformations. In this case it may be necessary to transform it into a series of period-to-period mean values. The new series shows an apparent linear upward trend as reveals the diagram from Fig. 4. The linear correlations of the data presented in Fig. 4 are:

- for period mean concentration: $c = 1.83 + 0.05\ t$
- for standard deviation: $\sigma = 0.321 + 0.009\ t$.

In both cases, the coefficients that reveal the time variation are very small and therefore the new TS can be considered stationary.

2.2 Second step of Box - Jenkins methodology

The second step in Box - Jenkins methodology is data smoothing with moving average method. This method formed new TS where each terms are an average of artificial subgroups created from consecutive observations. In Figs. 5 and 6 the smoothed data are the average of 3, 6 and 12 consecutive terms. The increasing subgroup data number, clearly put in evidence the daily periodical evolution of CO concentration during the considered period. The best results are obtained when $n = 12$.

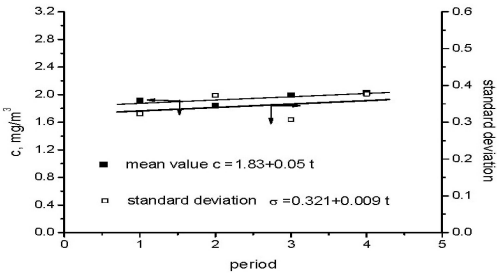

Fig. 4. The new TS with mean period values as terms.

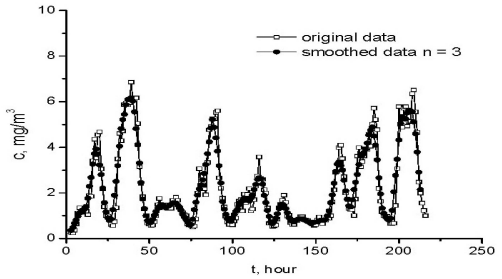

Fig. 5. Original and smoothed data of CO concentration in air on Bucharest, Mihai Bravu intersection during 1 - 9 March 2005; smoothed data were calculated with subgroups of 3 consecutive data.

The same results are obtained when are smoothed the daily and monthly average concentrations. The data presented in Figs. 7 and 8 reveal the weekly and seasonal sinusoidal variation.

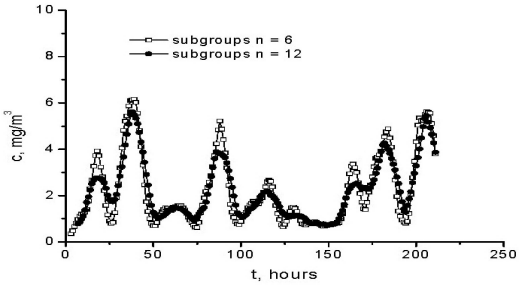

Fig. 6. CO concentration in air on Bucharest, Mihai Bravu intersection during 1 - 9 March 2005; smoothed data were calculated with subgroups 6 and 12 consecutive data.

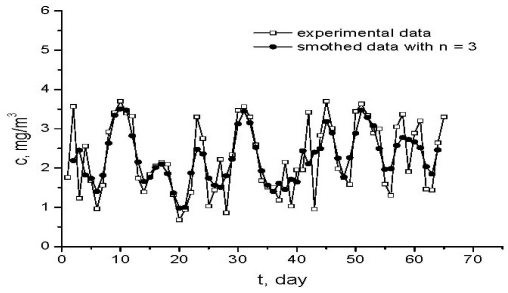

Fig. 7. Comparison between daily average CO concentration in the air and smoothed data, in Bucharest, Mihai Bravu intersection during March - April 2005.

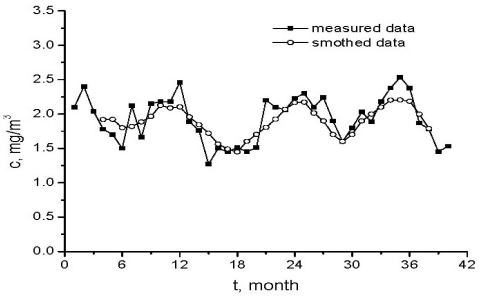

Fig. 8. Comparison between monthly average CO concentration in the air and smoothed data in Bucharest, Mihai Bravu intersection during January 2005 and June 2008.

2.3 Third step of Box - Jenkins methodology

The third step in Box - Jenkins methodology is to correlate the smoothed data in a mathematical model. The smoothed data from Fig. 6 shows sinusoidal variations of hourly CO concentration during days. The sine wave model can fit these data. The model constants are determined by nonlinear regression and the following relationship results:

$$c = 2.09 + 1.36 \cdot \sin\left(\frac{2\pi t}{24} + 3.65\right).$$

It can be noticed, that the period of sinusoidal variations of daily concentrations measured is equal to the duration of a night-day alternate. The differences between smoothed data and those calculated with mathematical model (Fig. 9) demonstrates the influences of the random factors, like meteorological conditions, on the measured values. In diagram from Fig. 7 the TS data show a weekly periodicity. In this case, the smoothed data was fitted in the following model:

$$c = 2.05 + 0.68 \cdot \sin\left(\frac{2\pi t}{7} + 5.51\right).$$

The differences from the smoothed data and the model data are still significant, but quite reduced relative to the case presented in Fig. 10. This demonstrates that by averaging the daily concentrations, the influence of random factors was diminished and the cyclic character of data becomes more evident.

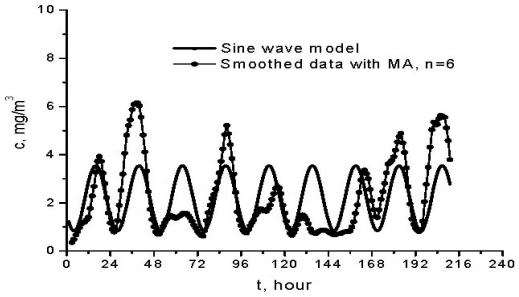

Fig. 9. Comparison between the sine wave model data and smoothed data of CO concentration in air in Bucharest, Mihai Bravu intersection during 1 - 9 March 2005.

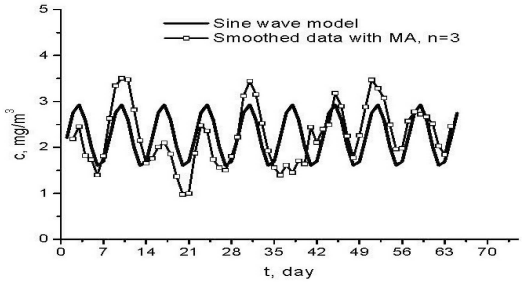

Fig. 10. Comparison between the sine wave model data and smoothed data of average daily CO concentration in the air in Bucharest, Mihai Bravu intersection during March - April 2005.

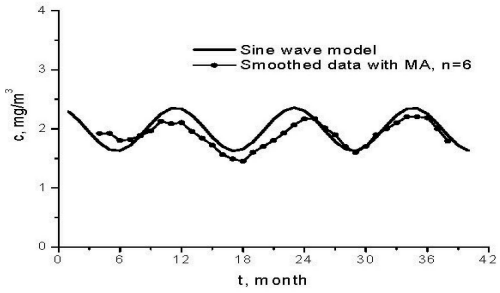

Fig. 11. Comparison between the sine wave model data and smoothed data of monthly average CO concentration in the air in Bucharest, Mihai Bravu intersection during January 2005 and June 2008.

In diagram from Fig. 11 the TS data show a seasonal periodicity. In this case, the deviations of experimental data from the smoothed data are still significant but quite reduced relative to the cases presented in Figs. 6 and 7. This demonstrates once again that by averaging the influence of random factors is diminished, while the periodical character becomes more evident.

The result of data correlation from Fig. 11 was the following equation:

$$c = 2.02 + 0.36 \cdot \sin\left(\frac{2\pi t}{12} + 1.64\right)$$

This equation reveals the seasonal variation of CO concentration in the air. In urban areas, the main source of CO pollution is the vehicular traffic producing emissions that are distributed all the year. The vehicles fuel consumption increases in cold weather, sometimes to 40% and the pollution outputs are much higher in the first minutes after a cold start of vehicles engines. The catalytic converter does not work when it is cold and the engine emissions pass through the exhaust untreated until the converter warms up and so results the monthly average variation of CO concentration that can be correlated with a *wave sin* time model with a period equal to 12.

The CO concentration in air will be dependent on the road traffic intensity (daily and weekly variations), technical state of vehicles, quality of the commercial fuels, but also the average month temperature during cold season and its duration (seasonal variation). The mathematical models obtained by nonlinear regression can provide the time evolution of a pollutant as long as the process is stationary.

3 Conclusions

The paper demonstrated that, in a mathematical approach, the experimental data obtained on monitoring the pollutant concentration can be considered as a time series. The application of time series properties to a set of experimental data obtained on monitoring CO concentration in air over a four years period in Bucharest Mihai Bravu intersection highlights the peculiarities of the time evolution of the monitored variable and provide useful information for further predictions of the running process. Statistical processing of TS elements requires validating the existence of the following properties: variability, homogeneity, periodicity, interdependence and stationarity. The first stage in Box - Jenkins methodology put in evidence these properties.

In the second stage, the TS data are smoothed in order to reduce the random fluctuations and put in evidence the trend of collected data.

The correlation of smoothed data with a *wave sine* model represented the third stage. The models reveal the daily, weekly, and seasonal periodicity of CO concentration in air.

References

[1] N. Barbalace, L. Giannetto, *Modeling of atmospheric pollutants in petrochemical refineries*, BSG Proc. 12, Geometry Balkan Press 2005, 23-31.

[2] A. de Brauwere, F. de Ridder, R. Pintelon, J. Meersmans, J. Schoukens, F. Dehairs, *Identification of a periodic time series from an environmental proxy record*, Computers & Geosciences, 34 (2008), 1781-1790.

[3] K. Hipel, A. I. Mcleod, *Time Series Modeling of Water Resources and Environmental Systems*, Elsevier, 1994.

[4] I. Meghea, I. Lacatusu, M. Mihai, I. Popa, *Air Pollution, Monitoring and Statistics of Air Pollutants*, Politehnica Press Eds., 2010.

[5] I. Meghea, M. Mihai, I, Lacatusu, T. Apostol, *Environmental monitoring of CO emissions: statistical character of acquired data*, EEMJ 8, 3 (2009), 575-582.

[6] I. Meghea, M. Mihai, E. Craciun, *Monitoring and statistics of heavy metals daily data in surface water*, SGEM 2010, Vol. II, 677-683.

[7] I. Meghea, M. Mihai, I. Popa, S. Ganatios, *Statistical analysis of hourly ground level of primary pollutants concentrations in Bucharest*, Proc. of RICCCCE XVI, 2009, 7-16.

[8] M. Mihai, I. Meghea, T. Necsoiu, A. Meghea, *Effect of thermal inversion on Bucharest air pollution*, Proc. of RICCCCE XVI, 2009, 17-26.

[9] I. Meghea, M. Mihai, I. Popa, M. E. Ganatiou, *Modelling study for forecasting of urban air pollution*, Proc. of Int. Symp. SIMI 2009 - "Environment and Industry", 2009, 31-43.

[10] G. Sakalauskiene, *The Hurst phenomenon in hydrology, environmental research, engineering and management*, 3 (25) (2003), 16-20.

[11] S. Sridharan, *Statistical properties of hyperbolic Julia sets*, Diff. Geom. Dyn. Syst. 11 (2009), 175-184.

[12] C. Tarcolea, A. Paris, A. Dumitrescu-Tarcolea, *Statistical methods applied for materials selection*, Appl. Sci. 11 (2009), 145-150.

[13] J. M. Zaldivar, F. Strozzi, S. Dueri, D. Marinov, J. P. Zbilut, *Characterization of regime shifts in environmental time series with recurrence quantification analysis*, Ecological Modeling 210 (2008), 58-70.

Authors' addresses:

Mihaela Mihai
University Politehnica of Bucharest,
Faculty of Applied Chemistry and Materials Science,
1 Polizu Street, 011061 Bucharest, Romania.
E-mail: mihai_mihaela2007@yahoo.com

Irina Meghea
University Politehnica of Bucharest,
Faculty of Applied Sciences, Department of Mathematics II,
313 Splaiul Independentei Street, 060042 Bucharest, Romania.
E-mail: i_meghea@yahoo.com