Journal of mathematics and computer science 7 (2013), 121-130

Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Trend Analysis with Effective Covariates Based On Auto Regressive-Moving Average Time Series Residuals

Manoochehr Babanezhad

Department of Statistics, Faculty of sciences, Golestan University,
Gorgan, Golestan,Iran

m.babanezhad@gu.ac.ir

Article history:
Received March 2013
Accepted April 2013
Available online April 2013

Abstract

Determining the pattern of a time series data is commonly established through identifying trend analysis. There is a variety of regression approaches can be chosen to perform trend analysis.All regression models are differentto the choose of which confounding factor are adjusted in the model.In view of this, when one takes into account the effective covariates in the trend analysis model,different patterns of a considered time series data setare created at each time t. This study proposes a methodology for characterizing the long term evolution of particular matterto identifyair quality analysis in the presence of radium, temperature and wind direction with correlated residuals in multiple regression models. Moreover, this is interesting in case where one performs trend analysis of the evolution of particular matters in quantity to air quality with significant effective covariates. Specifically, the considered approach provides a frame work based on the Gaussian correlated residuals where they follow a stationary Auto Regressive-Moving Average (ARMA) time series model.

Keywords: Time Series, Trend, Covariates, Particular Matter, Residuals, ARMA model.

1. Introduction

The long term evolution of particular matter (PM_{10}) to identify air quality patterns is recently a subject of growing interest [1, 2]. This may be because PM_{10} consists of tiny solid and liquid particles that come from a myriad of sources, both natural and human-caused. PM_{10}particles rangenormally lies in size over many orders of magnitude. They areoften divided into a fine fraction (less than 2.5 micrometres, called "$PM_{2.5}$") and a coarse fraction (2.5-10 micrometres). PM_{10} indeed includes both $PM_{2.5}$ and coarse PM. When long term evolution of PM_{10} is varying in a period of time t, determining the pattern of this evolution is commonly important to establish of air pollution of an urban even rural areas. In view of this, it is well known that air pollution problem affects not only the big cities but also medium sized urban areas [2, 3, 4]. A typical example that is considered in this study is a small city with a very small population in the province of Tehran, Iran which is located in the continent/region of Asia. Recent studies have revealed that the air pollution problem in these areas is caused due to increasing radium andtemperature. This might be due to the fast industrial growth in the last decades that resulted in an increase of the sources of pollution. Further, this sudden growth in combination with the urban development has resulted to increased traffic throughout the central district of the city, constituting direct sources of PM pollution and causing re-suspension of dust. In addition, particulate pollution is of paramount importance in areas with open-pit mines and especially when it is combined with raw lignite transfer and combustion in power stations (PS) through the suspension of particles and stack emissions, respectively. To grad with trend effect variation, one often employees a systematic class of multiple (linear or nonlinear) regression models. The approach includes a provision for treating on the different degrees of time. The defining feature of these models is that they are multiplicative models, meaning that the observed data are assumed to result from time effect. In view of this, this paper has two goals. The first aim is to show the trend analysis under the linear or nonlinear regression models leads to the biased estimation where it merely relies on the polynomial of the time effect in the multiplicative regression model [5, 6, 7]. This setting seems reasonable in practical situations in which trends are expected to change almost surly at the different levels of effective covariates [7, 8, 9, 10]. More specifically, this article deals with the situation in which an explanatory variable (orvariables) of interest evolves over the time t, and is measured at several different points in timeon each of a number of units (or subjects). Interest lies in controlling the effect of this time varying explanatory variables (or time varying confounders).Failure to do so typically results in a biased estimator of the trend effect of interest. This paper focusesin fact on the specificproblem of timevarying confounders, i.e. factors that potentially confound the trend over the time andare measured repeatedly throughout the study.

The parameter estimation is often performedby ordinary maximumlikelihood (For more details see[5, 6]). [5, 6] statethat, for moderate sample lengths, the EML estimator is generally less biased than the other estimators and leads to more accurate inference provided that the residuals (error terms) of the regression model to be uncorrelated. This assumption however is untestable assumption, but it is more important to control the behavior of error terms in the considered regression models. Because the validation analysis aims to check if the assumptions of themodel (the relationship between the response and the covariates,the time series models structure and the normality of the error terms) are correct since, otherwise, the conclusions obtained from the modelcould not be true.

Second goal is to develop statistical procedures to test whether the trend analysis model has correlated residuals or not. Specifically, we show that ordinary regression model is not capable of dealing with the air quality analysis when correlated time varying residuals may follow a stationary ARMA model.A substantively different, yet methodologically closely related, problem arises when we wish to analyze thetrend effect of a time series, acting through some mediator variables where they may be directly affect pattern of trend of considered time series data set.

2. Trend Component

An important step in analyzing time series data is to consider the types of data patterns, so that the models most appropriate to those patterns can be utilized. One important type of time series components is trend which is in fact long term increase or decrease in the data [1, 11, 12]. In other words trend is indeed the component of a time series that describes long term variation of mean in a time series data setting. These variations can be viewed of low, high, or of medium frequency fluctuations having been filtered out. There is a variety of multiple regression approaches can be chosen to identify trend analysis. However an important problem in estimating trend effects is to adjust all effective covariates which directly or indirectly affect the response variable in the multiple regression model. For doing so, the procedure involves a distribution-free estimate of trend in multiple linear or nonlinear regression models. Suppose that sequence of random variable $\{X_t\}$ is a time series and the data $X_1, \ldots, X_n$ are observed from this with the following model [5, 13];

$$X_t = g(t; \beta) + e_t \quad (1)$$

for $t = 1, \ldots, n$, where g(.) is an unknown regression function, and error time series $\{e_t\}$ is white noise random process with mean 0 and with a constant variance. The process X_t is mean non-stationary and can be interpreted as a signal g(.) plus noise et model. Specifically $g(t; \beta)$ represents the time evolution, modeled by a polynomial, e.g., $g(t; \beta)=\beta_0+\beta_{1t}+.....+\beta_{tk}$, whose order is determined during the modeling process. Note that the error time series $\{e_t\}$ may be to have Gaussian distribution.The parameters β_0, ,$\beta_1,...,\beta_k$ are unknown and should be estimated. The parameters estimation is performed by restricted maximum likelihood estimation (RMLE) [5, 6, 14]. However ordinary maximum likelihood estimation (MLE) could also be applied but for moderate sample length, RMLE estimator is, generally, less biased than the MLE and leads to more accurate inference [5, 7, 15, 16]. The power of t in regression model is mainly based on the values of the probability levels. The model selection is carried out in a systematic and iterative way, starting at the simplest model and then comparing models in a stepwise approach, until obtaining satisfactory fitting with data.

3. Particulate Matter Data

Particulate pollution can harm the human respiratory and cardiovascular systems, and is linked to asthma and mortality. Smaller particles are the most damaging and current targets focus on particles less than 10μm in diameter (PM_{10}). Coal burning, diesel combustion, construction, mining and quarrying are the major sources of particulateemissions. Atmospheric particulate matter - also known as particulates or particulate matter (PM10) - are tiny pieces of solid or liquid matter associated with the Earth's atmosphere. They are suspended in the atmosphere as atmospheric aerosol, a term which refers to the particulate/air mixture, as opposed to the particulate matter alone. However, it is common to use the term aerosol to refer to the particulate component alone. Furthermore, the smaller and lighter a particle is, the longer it will stay in the air. Larger particles (greater than 10 micrometers in diameter) tend to settle to the ground by gravity in a matter of hours whereas the smallest particles. In this study we consider a small city with a very small population in the province of Tehran, Iran which is located in the continent/region of Asia.The data variability is large in January 1995 through March 2004 (Figure 1). In addition to PM_{10}, there are many factors affecting such as radium, temperature, and wind direction. The statistics trend analysis of the data setting is preformed in the next sub sections.

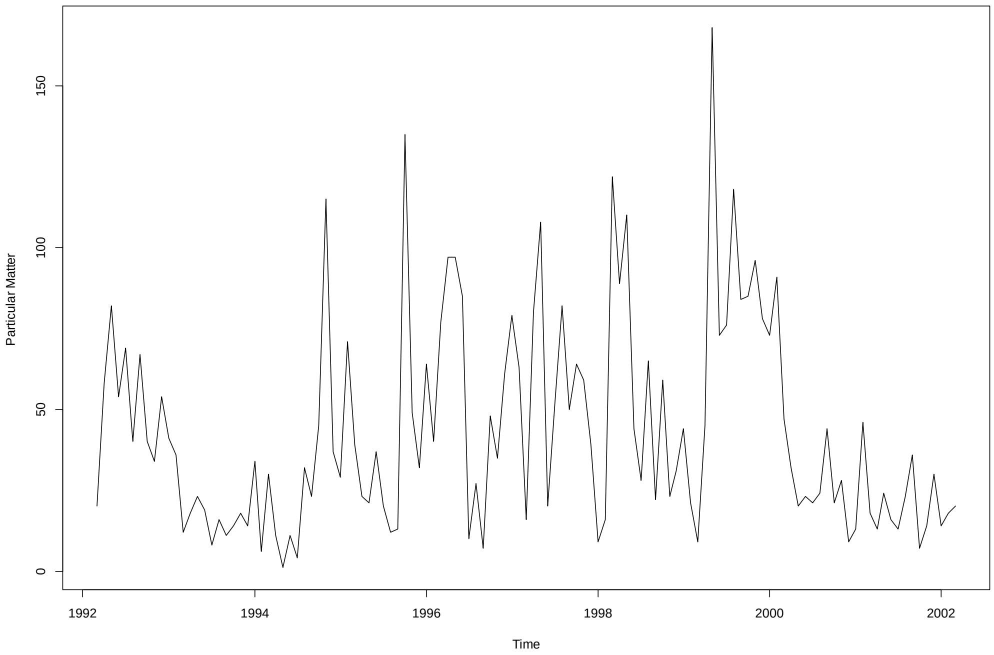

Figure 1. Yearly time series of conductivity (points) and long term evolution of PM_{10} January 1995 through March 2004.

3.1. Trend Analysis

The classical descriptive analysis is the first statistical analysis dealing with any data. Mean, standard deviation (STDEV), maximum and minimum value of selected data sets are usually calculated to have preliminary knowledge of selected variables. The calculation of coefficient of variation (CV) helps the investigator to overcome the problem of different levels and units of variables in order to compare them. We then draw time series plot. Time series plot can present a preliminary understating of the time behavior of the series. Fig.1. shows time series plot of selected time series air pollution concentration. This Figure shows different time behavior of PM_{10}. Trend monitoring looks for changes in environmental parameters over time periods. The autocorrelation functions of the selected time series also show different time stationary of the series. The autocorrelation functions of them are presented in Fig. 2. In this section, we perform trend analysis of PM_{10} data in two cases. The model fitted (2) produces a value R^2=0.20 which is the square of the residuals of the data after the fit. It says what fraction of the variance of the data is explained by the fitted trend line. It does not related to the statistical significance of the trend line indeed. Often, filtering a series increases R^2 while making little difference to the fitted trend.

3.1.1. Trend Analysis without Covariates

To investigate whether PM_{10} follows a pattern of time dependent, we regress PM_{10} on the polynomial of time effect in nonlinear regression models [9, 10]. The polynomial time regression between PM_{10t} and time is postulated as follows:

$$PM_{10t}=\beta_0+\beta_1 t+ \beta_2 t^2 + \beta_3 t^3+ \beta_4 t^4+ e_t \qquad (2)$$

The model (2) supposes that there are two components of variability for the PM_{10} at each time t; the mean value varies with time and the difference from the mean varies randomly. Time is the only factor affecting the mean value, while all other factors are subsumed in the noise component. In equation (2), standard least squares regression could then be used to fit a linear model, and make predictions for PM_{10t}. Table 1 shows the effect estimated of linear regression model (2) in trend analysis when the time effect is merely adjusted in the regression model. The problem of time series analysis is to find the best form of the model for a particular situation. One problem here is when the effective covariates are excluded in trend regression model; the effects estimated may yield to be biased. That is covariates should be included in regression analyses if they are correlated with the PM_{10}. This is more important when the covariates are serially correlated with the PM_{10} at each time t. In the considered data, radium, temperature and wind direction are serially correlated with the PM_{10} at each time t. These interactive effect between the covariates and ambient particulate matter on mortality has attracted attention world-wide, but the results of studies investigating this interaction have been inconsistent. We found different patterns in relation to the effect of particular matter when these effective covariates are adjusted in the regression model.

Table 1. *PM_{10}trend analysis where merely the different degrees of time effects are adjusted*

Parameters	Estimate	S.E.	P.Values
β_1	-4.3600	1.56000	0.030
β_2	0.1500	0.06000	0.010
β_3	-0.0010	0.0060	0.001
β_4	0.0005	0.0002	0.040

R^2=0.20, MSE=843

3.1.2. Trend Analysis with Covariates

The model fitted (2) produces a value R^2=0.20 which is the square of the residuals of the data after the fit. It says what fraction of the variance of the data is explained by the fitted trend line. It does not related to the statistical significance of the trend line indeed. Often, filtering a series increases R^2 while making little difference to the fitted trend. Therefore radium, temperature and wind direction with different degrees of time effect are adjusted to the regression model as follows;

$$PM_{10t}=\beta_0+\beta_1 t+ \beta_2 t^2 + \beta_3 t^3+ \beta_4 t^4++ \beta_5\,\text{radium}+ \beta_6\,\text{temperature}+ \beta_7\,\text{wind}_+ e_t \qquad (3)$$

By performing step-wise algorithm selection to analysis the latter model, the significant effect of radium, temperature, and wind direction were observed. The analysis is summarized in Table 2.

Table 2.*PM_{10} trend analysis where the different degrees of time effects and time varying covariates are adjusted*

Parameters	Estimate	S.E.	P.Values
β_1	-1.20	0.580	0.020
β_2	0.190	0.010	0.010
β_3	-0.001	0.007	0.020
β_5	0.460	0.240	0.001
β_6	1.820	0.330	0.002
β_7	-2.560	0.063	0.003

R^2=0.67, MSE=136

Although this approach controls for time-invariant confounders by design, it may allow for selection bias and confounding by time-varying factors. Schwartz and Marcus (1995) [3] demonstrated that confounding by seasonal and long-term time trends in environmental time-series data could be largely controlled by using the multivariate time varying regression models. We have examined that work to consider confounding by patterns with less than a year-long period. Here, we have induced patterns in the outcome owing to an omitted covariate that varies in conjunction with the pattern in exposure.

4. Estimating ARMA Residuals Structure

For the second goal of constructing a time series model for residuals $\{e_t\}$ in model (1), we claim that in all the PM_{10} series of this study, residuals correlation was detected after fitting regression model (1). Time series regression usually differs from a standard regression analysis, because the residuals form a time series and therefore tend to be serially correlated. Whether this correlation to be positive or negative, the estimated standard errors of the parameter estimated may lead to be less or more than their true value. In this study, an interesting problem is to test whether the residuals of the trend analysis regression model are correlated, and they follow a time series models. This can be checked through Auto-correlation and partial Auto-correlation. The general model introduced by Box and Jenkins (1976) [7] includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters (q). In the notation introduced by Box and Jenkins (1976), models are summarized as ARIMA (p, d, q); so, for example, a model described as (1, 2, 2) means that it contains 1 order autoregressive parameters and 2 moving average parameters which were computed for the series after it was differenced two times. Differencing turns out to be a useful 'filtering' procedurein the studyof non-stationary time series. By plotting Auto-correlation function (ACF)(Figure 2; right panel) and Partial Auto-correlation function (PCAF) of second order of differencing {et}; that is $w_t=e_t-e_{t-2}$ (Figure 2; left panel), we realized that CAF and PCAF suggest that $\{w_t\}$ in model (3) follows an autoregressive and moving average process model of order (1, 2) denoted with ARMA(1,2) as follows;

$$w_t=\alpha_1 w_{t-1}+ z_t+\alpha_2 z_{t-1}+ \alpha_3 z_{t-2} \qquad (4)$$

where {zt} is white noise time series with mean 0, and constant variance, and the parametersα_1,α_2, α_3 are unknown and should be estimated.A summary of the main results of the latter model is provided in Table 3.

Table 3.*PM$_{10}$trend analysis where merely the different degrees of time effects are adjusted*

Parameters	Estimate	S.E.	P.Values
α_1	0.230	0.730	0.040
α_2	0.420	0.072	0.001
α_3	-0.080	0.020	0.003

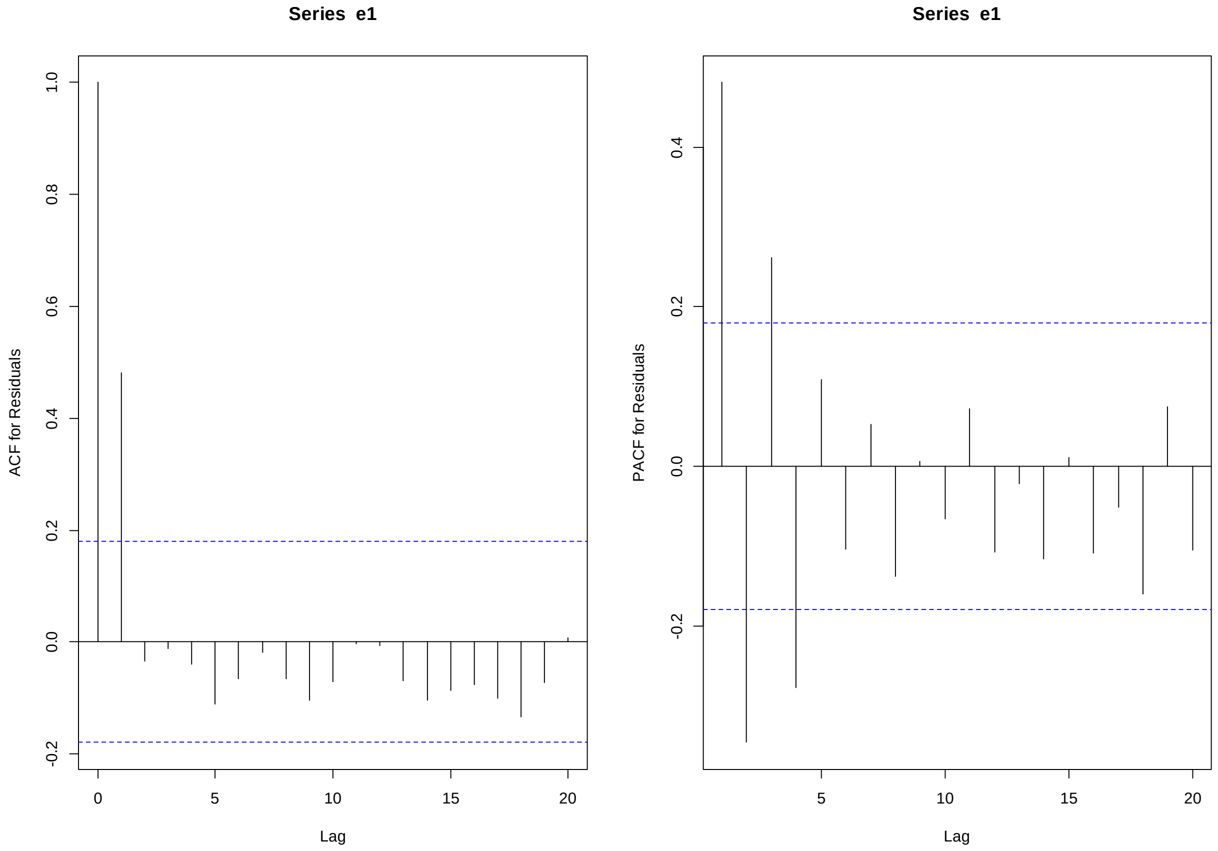

Figure 2. Correlograms of the ACF (right panel) and PACFleft panel) of filtered residuals from the second order of differencing {et}; $w_t=e_t-e_{t-2}$.

5. Discussion

Daily particular matter time series analysiswas performed in this study. The analysis showed different temporal behavior of different air pollutants. This study was in fact designed to analyze the effective of particular matter and radium, temperature and wind direction temperature on population mortality on identifying air quality in the considered city. To do so, we first control the pattern of trend which is called the potential pattern of over time: linear (where the mean is steadily increasing or decreasing over time), quadratic (wherethe mean first increases and then decreases over time, or the reverse), or something more complicated.The trend is a significant change over time exhibited bya random variable,detectable by statistical parametricand non-parametric procedures. In problem concerning trend analysis, we have reiterated that standardregression analyses are invalid when confounders affected the patterns of considered time series data sting.The parametric test considers the linear regression ofthe random variable *Y* on time *t*. The regressioncoefficients*(*or the Pearson correlation coefficient) are theinterpolated regression line slope coefficient computedfrom the data.As for the identification of time series changing points,a preliminary regression merely on time t is highly instructiveand meaningful.The magnitude of the trend is resulted invalid and not highly significant in particular time intervals of thereference period, as indicated in corresponding testresults (Table 1).

Standard inference procedures for regression analysis always adjust tovariables that are effective in practice. However adjustments must be made to insure the validity of statistical inference. These adjustments are sometimes used routinely toimprove prediction and inference based on regression analysis.Because confounding canreduced through matching in the study design but this can be difficult and/or wasteful of resources. Another possible approach—assuming data on the confounder(s) have been gathered—is to apply a statistical 'correction' method during analysis. Such methods produce 'adjusted' or 'corrected' estimates of the effect of exposure; in theory, these estimates are no longer biased by the erstwhile confounders. In this paper we tested the hypothesis that in addition to the some degrees of time effect, radium, temperature, and wind direction modified the effect of ambient PM_{10} by adding in multiple linear regression models. The results showed that high radium and high temperature can enhance the effect of PM_{10} (Table 2). However there is not an interactive effect between radium and temperature and PM_{10} concentration. While some pollutants like radiation and temperature show simple temporal fluctuation through the year, wind direction show high fluctuation and have mostly non linear trend through the year using time series regression. High coefficient of variation and kurtosis in most of the observed series also indicated non linearity variation of PM_{10}concentration at each time t. It is also found different patterns in relation to the effect of particular matter and time effects when the effect of radium, temperature, and wind direction were not included in the trend analysis models. Moreover, it is found a model with low R^2 adjusted and high mean square error. Several studies [1, 15] have explored whether temperature modifies the adverse effects of PM_{10} on air quality, but these have not shown consistent results. Some studies have found interaction effects of radium and wind direction with PM_{10} concentration. While in this study, these effects are not observed in the regression analysis. The discrepancies between these findings may be influenced by many factors, such as geographic conditions and population characteristics [14, 15]. The analytical methods used in these studies may also contribute to this lack of consistency. The selection method in this study has several advantages. First, a dose-response curve allows the visual determination of potential turning points. Second, R^2 adjusted, mean square error and AIC are sensitive enough to obtain the appropriate stratification points, allowing the determination of parametric estimates, which are simple to interpret. To conclude whether there exists a linear trend in the data sequence; one should first decide whether there is persistence in the data, which will decide how the Mann-Kendall test is to be performed. Moreover, the important problem that have not been yet addressed is that the residuals of selected time varying regression model follows a pattern that suggest ARMA model. The residuals analysis in section 4 shows that the ARMA regression is able to detect underlying positive temporal trend, after eliminating the initial trend effect by second difference of a time series which is the series of changes from one period to the second. The ARMA regression detects a temporal trend in years 1996 and 1998. While the trend analysis of the residuals of model (1) have not detected this behavior. Moreover the possible mistakes resulting from not accounting for the serial correlation are evident in the year 1996 and 1998, where the regression model found a statistically significant positive temporal trend. It can be finally concluded that the use of regression model with a Gussian ARMA error term is more effective than the models not take into account. Therefore, the use of a regression model with aGaussian ARMA error is a powerful tool to characterize, quantifyon the temporal evolution of$PM_{10}$and to make inference air qualityparameters, under the difficult conditions often found in these series. Although the modeling process can be more timeconsumingthan other simpler approaches, it has important advantages assummarized in this paper. As a result, this paper deals with the limitations of other common approaches toPM_{10}trend analysis. More precisely, it can be used in series with a serial correlation structure.Because the proposed approach can characterize complextemporal evolutions such as non-monotonic trends, it offersan advantage over non-parametric trend analysis.

6. ACKNOWLEDGEMENTS

This paper was granted by Golestan University. Therefore I would like to thank Golestan University to support this project.

7. References

[1] S. Roberts, ‘‘Combining data from multiple monitors in air pollution mortality time studies”, Journal of Atmos. Environ. 37,3317-3322**(2003).**

[2] C. K. Lee, D. S. Ho, C. Yu, Wang C. and Y. Hsiao, “Simple multifractal cascade model for air pollutant concentration time series”, J. Environmetrics., 14, 255-269**(2003).**

[3] J. Schwartzand A. Marcus,“Mortality and airpollution in London: a time series analysis”,A.J. Epidem., 131, 85-194 (**1995).**

[4] G. Touloumi., R. Atkinson and A. L. Terte, “Analysis of health outcome time series data in epidemiological studies”, Environmetrics, 15, 101-117 **(2004).**

[5] R. H. Shumway and D. S. Stoffer, “*Time Series Analysis and Its Applications: With R Examples*”, Springer **(2010).**

[6] R.O. Gilbert, “*Statistical Methods for Environmental Pollution Monitoring*”, New York **(1987).**

[7] G.E.P.Box and G.M. Jenkins, “*TimeSeries Analysis, Forecasting And Control*”, Revised edition, Holden-Day, San Francisco **(1976).**

[8] F. Dominici, A. McDermott, S. L. Zeger and J. M. Samet, “On the use of generalized additive models in time-series studies of air pollution and health”, American journal of epidemiology, 156, 193-203 **(2002).**

[9] B. L. Bowerman and R. T. O'Connell, "*Time Series Analysis Forecasting Statistical Methods* "**,** 3rd edition **,** Duxbury Press (Belmont, Calif.) **(1993).**

[10] G. Li, J. Sun, R. Jayasinghe, X. Pan, M. Zhou, X. Wang, Y. Cai, R. Sadler, G. Shaw,**"**Temperature modifies the effects of particulate matter on non-accidental mortality: A comparative study of Beijing, China and Brisbane, Australia",Public Health Research , 2, 21-27 **(2012).**

[11] M. Stafoggia, J. Schwartz, F. Forastiere, C. A. Perucci, **"**Does temperature modify the association between air pollution and mortality? A multicity case-crossover analysis in Italy**",**American Journal of Epidemiology**,** 167, 1476-85 **(2008)**.

[12] C.M. Wong, N. Vichit-Vadakan, H. Kan, Z. Qian Z, "Public health and air pollution in Asia (PAPA): A multicity study of short-term effects of air pollution on mortality", Environ Health Perspect., 116, 1195-202 (**2008).**

[13] Y. Guo, K. Punnasiri and S. Tong,"Effects of temperature on mortality in Chiang Mai city, Thailand: a time series study"**,**Environmental Health,http://www.ehjournal.net/content/11/1/36 **(2012).**

[14] W. Yu , P. Vaneckova, K. Mengersen, X. Pan and S.Tong,**"**Is the association between temperature and mortality modified by age, gender and socio-economic status?",Sci Total Environ., 408, 3513-8 **(2010).**

[15] F.Dominici, J. M.Samet, S. L.Zeger, “Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy”, Journal of the Royal Statistical Society: Series A (Statistics in Society), 163,263-302 **(2000).**

[16] H. Azami, M. Malekzadehand S.Sanei,‘‘ A new neural network approach for face recognition based on conjugate gradient algorithms and principal component analysis’’, Journal of mathematics and computer Science, 6, 166-175 **(2013).**