Application of Time Series Models in Chemical, Biological and Environmental Sciences

Neelam Rani
Assistant Professor
Department of Mathematics
Arya Kanya Mahavidyalya Shahabad, Distt. Kurukshetra, Haryana, India.

Abstract: An autoregressive integrated moving average (ARIMA) is one of the popular linear models in time series forecasting during the past three decades. Recently many environmental time series data can be adequately modelled using the seasonal ARIMA model. Different models for using time series approach such as Autoregressive (AR) models, Autoregressive Moving Average (ARMA) models, Autoregressive Integrated Moving Average (ARIMA) models are discussed. The methodology involves Initial Model Development Phase, Parameter Tuning Phase and Forecasting Phase.A seasonal ARIMA models applied on PM2.5 data extracted from the California Air Resource Board.The whole process carry the three step of Box Jenkin modelling identification, estimation and diagnostic.

***Key Words*:** ARIMA; Seasonal ARIMA; Forecasting; PM2.5

I. Introduction

Air pollution is an effected by many environmental factors, and the factors have a complicated correlation, its hard to explain the correlation with a structure casual model. At this time, it is an effective method to establish the time series dynamic model in accordance with its own law, and air pollution often have long-term trends, seasonal, cyclical, short-term fluctuations and irregular changes [1]. Particulate matter (PM) is a mixture of all microscopic solid and liquid particles, of human and natural origin, that remain suspended in a medium such as air for some time. These particles vary greatly in size, composition, and origin, and may be harmful. Particulate matter may be in the form of fly ash, soot, dust, fog, fumes etc. PM2.5 refers to particulate matter that is 2.5 micrometres or smaller in size. The sources of PM2.5 include fuel combustion from automobiles, power plants, wood burning, industrial processes, and diesel-powered vehicles such as buses and trucks. In this paper time series model has been applied on the PM-2.5 data, which was produced by the California Air Resources Board.

Lon-Mu and Chung [2] employed time series analysis, in impact studies of environmental data. An iterative procedure for the joint estimation of model parameters and outlier effects was employed with the intervention analysis. They found that this joint estimation procedure not only produces more reliable estimates of intervention effects, but also provided information on outliers, which was valuable in many respects.

Ynng et al. [3] applied the time-series analysis theory to analyse and forecast the dynamic variation of groundwater in west part of Jilin Province, China. First, the trend component of groundwater level dynamic variation was picked up by polynomial calibration, the periodic component was extracted by spectrum analysis and the stochastic component was simulated by using autoregression (AR) model. Finally, a forecasting model was established through linear superposition of these components and the method for verifying the accuracy of the model was suggested. The analysis of this model showed that there were two major periods in the variation of groundwater level in that area by which seasonal and secular variation of groundwater level was revealed. The prediction for years after 2002 indicated that a continuing decline of groundwater level exists in some district of that area, which must be controlled in time.

Ghil et al. [4] provided crucial information to describe, understand, and predict climatic variability for the analysis of univariate or multivariate time series. They reviewed connections between time series analysis and nonlinear dynamics. The various steps, as well as the advantages and disadvantages of these methods, were illustrated by their application to an important climatic time series, the Southern Oscillation Index. This index captured major features of interannual climate variability and was used extensively in its prediction.

The remainder of this paper is structured as follows: Section-2 introduces basic theory of time series forecasting. Section-3 introduces basic theory of ARIMA models. Section-4 reports each step presented to analyse time series data using SARIMA model. Finally, Section-5 presents conclusion of the study is discussed.

2 **Basic theory of time series forecasting**

What Is Forecasting? Process of predicting a future event. Some of the areas in which forecasting plays an important role:

- Scheduling
- Acquiring resources
- Determining resource requirements.

Although there are many different area requiring forecasts, the preceding three categories are typical of the short-, medium-, and long -term forecasting requirements of today's organizations. Forecasting situation vary widely in their time horizons, factors determining actual outcomes, type of data patterns and many other aspects [5].

2.1 **Categorization of forecasting methods**

The appropriate forecasting methods depend largely on what information are available.

1. Qualitative forecasting method: If little or no quantitative information is available, but sufficient qualitative knowledge exists, e.g., forecasting how a large increase in oil prices will affect the consumption of oil.
2. Unpredictable forecasting method: If little or no information is available, e.g., predicting the effects of interplanetary travel.
3. Quantitative forecasting method: If sufficient quantitative information is available. It can be applied when two conditions are satisfied:
   - Numerical information about the past is available.
   - It is reasonable to assume that some aspects of the past patterns will continue into the future and this is known as assumption of continuity [6].

2.1.1 **Type of quantitative methods**

An additional dimension for classifying quantitative forecasting methods is to consider the underlying model involved. There are two major types of forecasting models:

- Explanatory models: It assume that the variable to be forecasted exhibits an explanatory relationship with one or more independent variables. Regression model and multiple regression model are the example of explanatory models.
- Time series models: Here, prediction of the future value is based on the past values of the same variable and / or past errors, but not on the explanatory variables which may affect the system. The objective of these models is to discover the pattern in the historical data series and extrapolate that pattern into future. Time series models used for forecasting include decomposition models, exponential smoothing models and ARIMA models [5].

2.2 **Time series**

A time series is a sequence of observations taken sequentially in time. A time series is denoted by $\{Y_t \mid t = 1,2,...,k\}$. Time series data can exhibit a variety of patterns: horizontal, seasonal, cyclical, trend and irregular. Many data series include combinations of the preceding patterns. General mathematical representation of decomposition approach is:

$$Y_t = f(S_t,T_t,E_t). \tag{2.1}$$

- Deterministic time series: If future values of a time series are exactly determined by some mathematical function such as

  $$Y_t = \cos(2\pi ft), \tag{2.2}$$

  the time series is said to be deterministic.
- Non-deterministic/Statistical time series: If future values can be described only in term of a probability distribution, the time series is said to be non-deterministic /statistical time series.

2.2.1 **Stochastic process**

A statistical phenomenon that evolves in time according to the laws of probability is called a stochastic process.

1. Stationary stochastic process: A very special class of stochastic processes, called stationary processes, is based on the assumption that the process is in a particular state of statistical equilibrium.
2. Strictly stationary stochastic process: A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin i.e., if the joint probability distribution associated with m observations Y_{t1}, Y_{t2}, ... ,Y_{tm}, made at any set of times t_1, t_2, ..., t_m, is the same as that associated with m observations Y_{t1+k}, Y_{t2+k}, ..., Y_{tm+k}, made at any set of times $t_1 + k$, $t_2 + k$, ..., $t_m + k$.
3. Non-stationary stochastic process: A stochastic process, which is not in the state statistical equilibrium i.e., the mean wanders (changes over time), and the variance (or standard deviation) is not reasonably constant over time.

2.3 **The basic steps in a forecasting task**

General guidelines which should be helpful in selecting the adequate model for a given data set have been explained in this section. The methodology to be followed to achieve the desired objectives includes following stages:

2.3.1 **Stages of forecasting**

1. Stationarity testing
   - Removing exponential trends.
   - Identifying polynomial and seasonal trends.
   - Testing for data stationarity through unit root testing.
   - Stationarizing time series through differencing.
2. Model order selection and identification
   - Computing autocorrelation and partial autocorrelation (ACF AND PACF).
   - Selecting models using formal tests: Akaike's information criterion (AIC).
3. Parameter estimation
   - Selecting candidate ARIMA and GARCH models for a given data set.
   - Creating and fitting time series models to a data set.

4. Diagnostic analysis
   - Testing residuals for normality.
   - Analyzing model diagnostics.
   - Inferring model residuals.
5. Forecast simulation
   - Forecasting data using fitted models.
   - Using in-sample forecasts to evaluate model appropriateness.

3 Linear filter model

The white noise a_t (a sequence of uncorrelated random variables with mean zero and constant variance, $E[a_t] = 0$, $var[a_t] = \sigma^2$) is supposed transformed to the process Y_t, by what is called a linear filter, as shown in Figure (1). The linear filtering operation simply takes a weighted sum of previous random shocks a_t, so that

$$Y_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots \quad (3.1)$$

$$Y_t = \mu + \psi(B)\, a_t, \quad (3.2)$$

where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$ is a linear operator. If the sequence ψ_1, ψ_2, is finite or infinite and absolutely summable, the filter is said to be stable and the process Y_t is stationary. The parameter μ is then mean about which the process varies. Otherwise, Y_t is non-stationary [7].

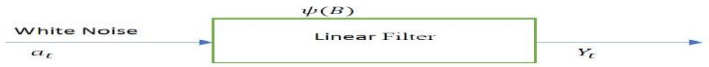


Figure 1: Representation of time series as the output from a linear filter

Depending on the characteristic of the linear filter, different models can be classified as follows:

3.1 **Autoregressive (AR) model**

In this model, the current value of the process is expressed as a finite, linear aggregate of the previous values of the process and the random shock a_t. Mathematically, an autoregressive (AR) process of order p for series Y_t is given below:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi Y_{t-p} + a_t. \quad (3.3)$$

If we define an autoregressive operator of order p in term of backward shift operator B by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p, \quad (3.4)$$

the autoregressive model (3.3) written as

$$\varphi(B)\, Y_t = c + a_t \quad [5]. \quad (3.5)$$

This model contains $p + 2$ unknown parameters $c, \phi_1, \phi_2, \ldots, \phi_p, \sigma_a^2$.
Autoregressive model is the special case of the linear filter model.

3.2 **The moving-average (MA) model**

In this model, we take Y_t linearly dependent on a finite number q of previous a's. Thus

$$Y_t = c + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\, a_t \quad (3.6)$$

$$Y_t = c + \theta(B) a_t, \quad (3.7)$$

where c = constant term, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$ is a moving average operator of order q. The model (3.7) contains $q + 2$ unknown parameters $c, \theta_1, \theta_2, \ldots, \theta_q, \sigma_a^2$.

3.3 **The autoregressive moving-average (ARMA) model**

To achieve greater flexibility in fitting of the actual time series, it is sometimes advantageous to include AR and MA term in the model.

$$\varphi(B) Y_t = c + \theta(B)\, a_t. \quad (3.8)$$

The model (3.8) contains $p + q + 2$ unknown parameters $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma_a^2$.

3.4 **The autoregressive integrated moving-average (ARIMA) model**

The ARMA models, described above can only be used for stationary time series data. However, load time series show non-stationary behaviour. In ARIMA models a non-stationary time series is made stationary by applying finite differencing of order (d) of the data points. The mathematical formulation of the ARIMA (p, d, q) model is given below:

$$\varphi(B)\,(1 - B)^d\,(Y_t - \mu) = \theta(B)\, a_t, \quad (3.9)$$

where backward difference operator is $\nabla = 1 - B$ and d is the number of regular differences.

3.5 **Seasonal autoregressive integrated moving average (SARIMA) model**

The seasonal difference operator is $\nabla_s = 1-B^s$, where s is the period of the seasonal cycle. Seasonal differencing will remove seasonality in the load series in the same way that ordinary differencing will remove a polynomial trend. A time series $\{Y_t \mid t = 1,2..., k\}$ is generated by SARIMA $(p, q, r)\ (P, D, Q)_s$ process with mean μ of Box and Jenkin time series model if

$$\varphi(B)\ \Phi(B^s)\ (1-B\ )^d\ (1-B^s)^D\ (Y_t-\mu) \quad = \quad \theta(B)\Theta(B^s)a_t, \qquad (3.10)$$

where p, d, q, P, D, Q are integer; $\varphi(B) = 1-\varphi_1 B-\varphi_2 B^2...-\varphi_p B^p$, $\Phi(B^s) = 1-\Phi_1 B^s-\Phi_2 B^{2s}-...-\Phi_P B^{Ps}$, $\theta(B) = 1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q$, $\Theta(B^s) = 1-\Theta_1 B^s-\Theta_2 B^{2s}-...-\Theta_Q B^{Qs}$, are polynomial of degree p, q, P and Q. B is the backward shift operator, D is the number of seasonal differences. The a_t should be independently and identically distributed (iid). The roots of $\varphi(B)$ and $\theta(B)$ should all lie outside the unit circle [8].

4 **Data Processing and implementation**

A monthly average is calculated from January 1999 to January 2006.The data is organized in a data frame, with value as rows and days as columns and the data is 12 x 8, with no missing data, so we have 96 observations for each variable. Time series data are divided into training data (84) and test data (12). The training data were used for estimating the model and the test data were used to measure the resulting model's error. The data are plotted as shown in Figure 2.

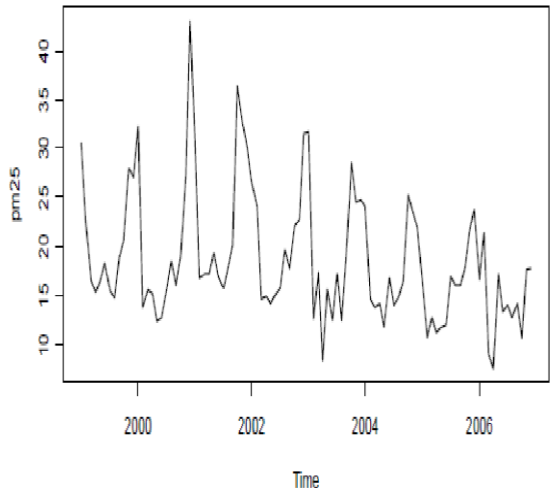


Figure 2: PM (particulate matter)2.5-month Average Concentration

Figure 2. shows the monthly PM 2.5 concentration levels from January 1999 through December 2006. There is obviously, a strong downward trend. Clearly, we need at least one order of differences.

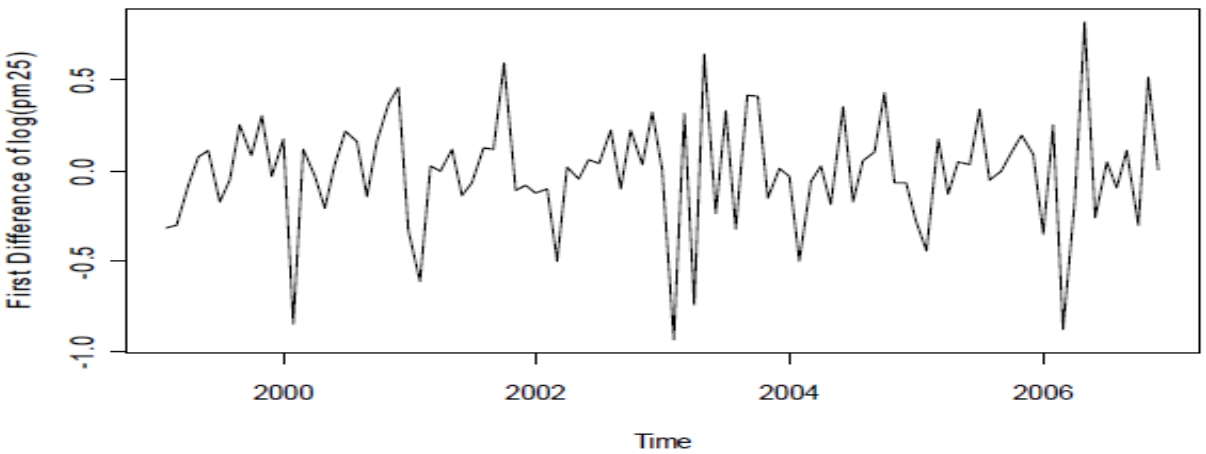


Figure 3: First Difference of Log (PM2.5) Levels

Figure 3: shows the time series plot of the log (PM2.5) levels after we take a first difference. The downward trend has now disappeared. The numbers of AR and MA can be identified in a more systematic way by ACF and PACF plot.

The fitted model is the following

$$(1 - B)\,(1 - B^{12)\,Yt} \quad = \quad (1 + \theta_1 B)\,(1 + \Theta_1 B^{12})\,a_t \tag{4.1}$$

$$Y_t \quad = \quad Y_{t-1} + Y_{t-12} - Y_{t-13} + a_t + \theta_1\, a_{t-1} + \Theta_1\, a_{t-12} + \theta_1\,\Theta_1\, a_{t-13}. \tag{4.2}$$

Table 1: Parameter Estimation

Coefficient	VALUE	SE
θ_1	0.8645	0.0790
Θ_1	0.6894	0.1268
AIC	495.3	

Thus, our fitted SARIMA $(0,1,1)\,(0,1,1)_{12}$ model is:

$$Y_t \quad = \quad Y_{t-1} + Y_{t-12} - Y_{t-13} + a_t + .8645\; a_{t-1} + .6894 a_{t-12} + .8645 * .6894\; a_{t-13}. \tag{4.3}$$

Diagnostic checking is necessary to ensure the best forecasting model has been built. The estimated SARIMA $(0,1,1)\,(0,1,1)_{12}$ model, provide the residual. Figure4: gives the time series plot of the residuals. The residuals even balance out around zero, it seems constant. And neither additive outliers nor innovative outliers have been detected.

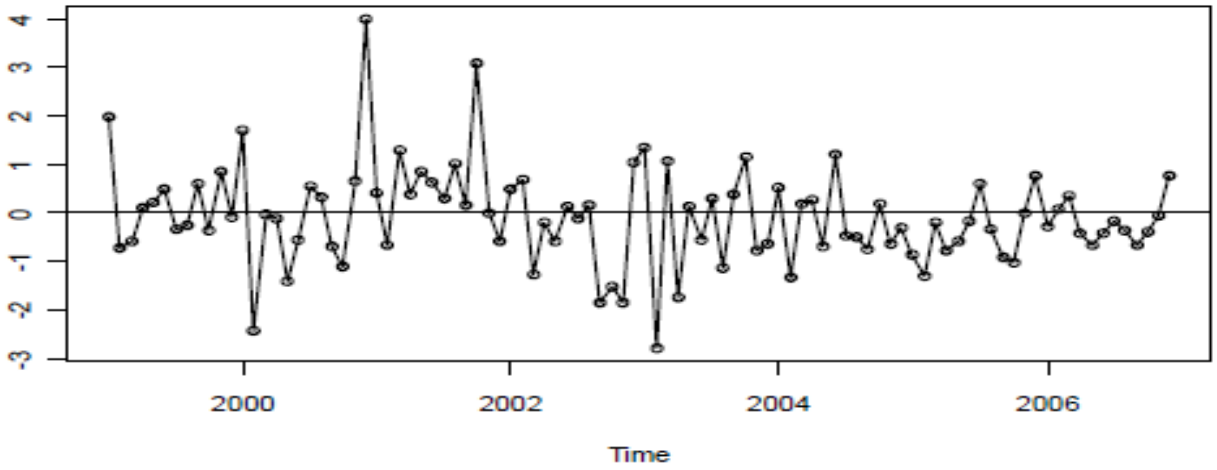
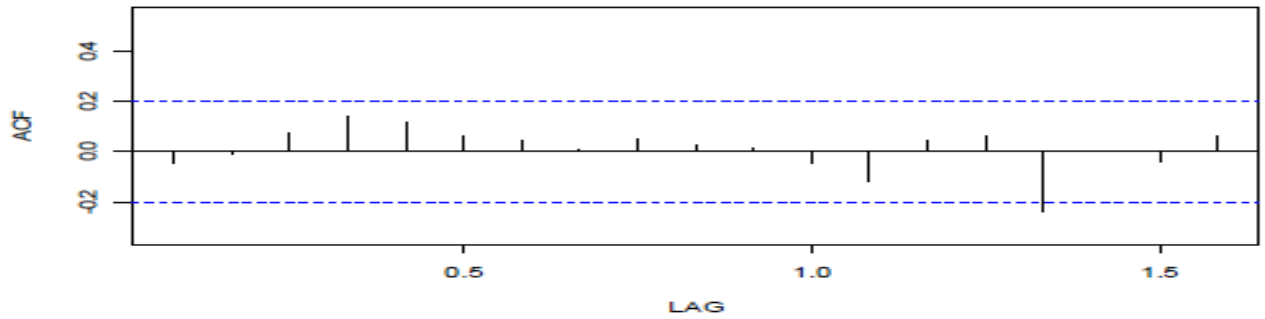


Figure 4: time series plot of the residuals.

To take a further look, we have plotted the sample ACF of the residuals. The sample autocorrelation function (ACF) plot with the vertical axis is almost close to 0 as in figure 5.



From the histogram (figure 6) of the residual it can be easily seen that the center seems a little bit off from zero, and the shape of histogram appears good shaped.

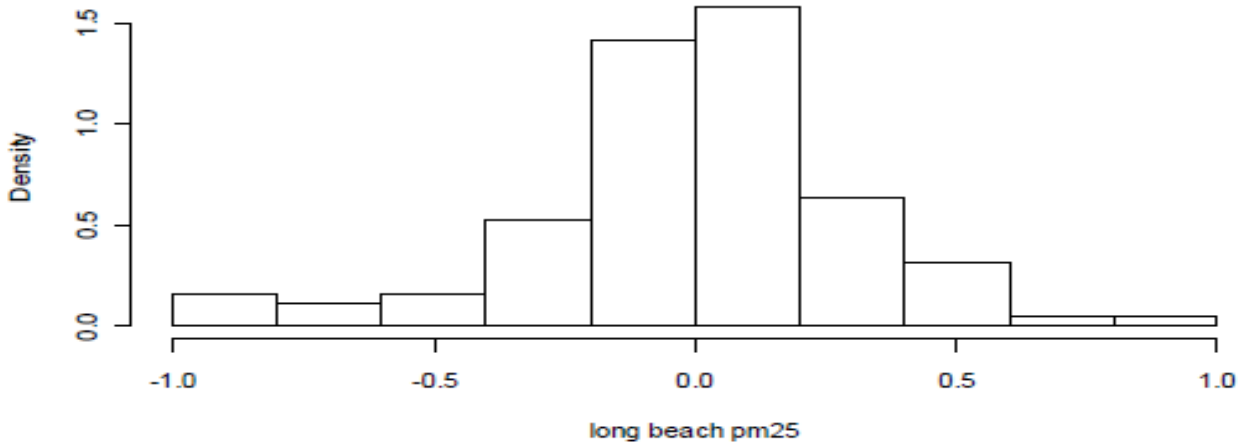


Figure 6: Histogram of the residuals

5. Forecasting Result

Figure 7: shows the forecasts and prediction interval with 95percent limits for a led time of one year for the SARIMA (0,1,1) $(0,1,1)_{12}$.

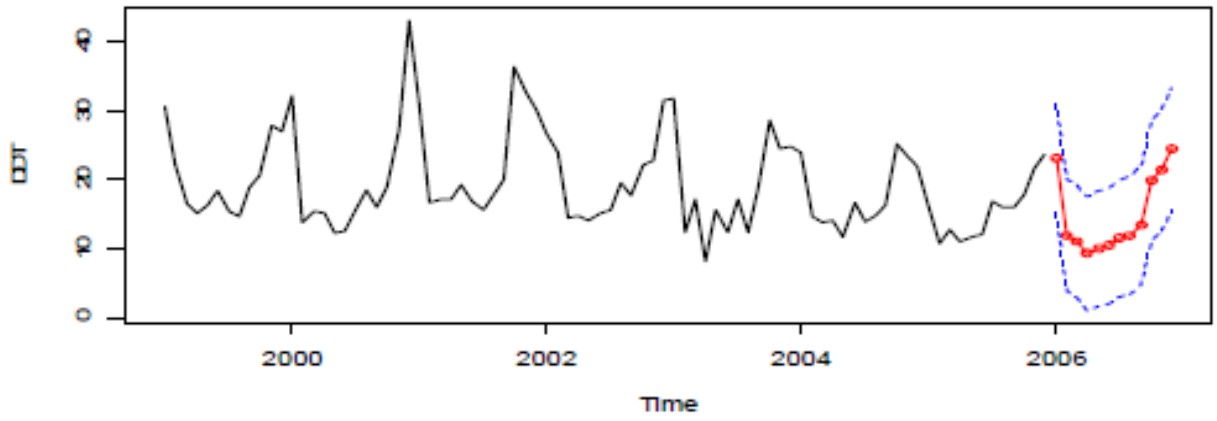


Figure 7: Comparison: Observed value and Predicted Value

The last one year of observed data is also shown. The forecasts mimic the stochastic periodicity in the data quite well, and the forecast limit give a good feeling for the precision of the forecasts.

6. Conclusion

In this study, SARIMA model was applied to predict the existing trend and seasonality in the PM2.5 data. The forecasting model was developed using ARIMA method in MATLAB 2017a software. The results indicate that *SARIMA* (0,1,1) $(0,1,1)_{12}$ is a suitable model based on analyzed historical data. The forecast of PM2.5 illustrates the pattern as well as the seasonality of the data. The model will be helpful to predict the air pollution PM2.5. For further work, it is still needed to evaluate and compare with other forecasting methods to obtain a better accuracy of forecast value.

References

[1] A. M. K.W. Hipel, Time Series Modelling of Water Resources and Environmental Systems, Elsevier, 1994.

[2] L. Lon-Mu, C. Chung, Recent developments of time series analysis in environmental impact studies, journal of environmental science and health 26 (7) (1991) 1217–1252.

[3] Z.-P. Yang, W.-X. Lu, P. Li, Application of time-series model to predict groundwater regime., Shuili Xuebao (Journal of Hydraulic Engineering) 36 (12) (2005) 1475–1479.

[4] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou, Advanced spectral methods for climatic time series, Reviews of Geophysics 40 (1) (2002) 3–1–3–41.URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000RG000092

[5] S. Makridakis, S. C. Wheelwright, R. J. Hyndman, Forecasting method and application, 3rd Edition, Wiley, 2011.

[6] R. J. Hyndman, G. Athanasopoulos, Forecasting: Principles and practice (May 2018).
URL https://otexts.org/fpp2/data-methods.html

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, Time Series Analysis Forecasting and Control, third ed. Edition, Englewood Cliffs, NJ: Prentice Hall, 1994.

[8] F. M. Tseng, H. C. Yu, G. H. Tzeng, Combining neural network model with seasonal time series arima model, Technological Forecasting and Social Change 69 (1) (2002) 71–87.