International Journal of Intelligent Computing and Technology
Vol. (1), No. (1), 2017, pp. 20-31

Analysis of Time Series for Modeling ARIMA in Household Annual Income

Dr.S.Prakashkumar[1] & K.Muthuchamy[2]
Assistant Professor[1,2] , Department of Computer Science
Theivaniammal College Arts and Science for Women
Villupuram, TamilNadu, India
drsp1974@gmail.com[1], muthuchamyka@gmail.com[2]

ABSTRACT

Household Annual Income as a case to illustrate how data mining can be applied to such time series. In this paper, an attempt is made to demonstrate the time series techniques in the context of data mining using freely available software MS-Excel with add-in XLMiner by providing a numerical illustration. Fitted the ARIMA models with various value of p, d, q parameters. For that, first, partition the given data into training data and validation data. Then, construct explorative techniques ACF and PACF for the both data and identify the tentative model. Fit a tentative model, and check whether the fitted model is adequate or not through results, ACF and PACF plots for residuals. The computed constant and the coefficient term for the model ARIMA (1,1,0) by using XLMiner and fitted forecasting equation is $Y_t = 3.825 + 0.978 Y_{t-1} + \varepsilon_t$. *It is observed that, for the non-seasonal data all the models except ARIMA (1,1,0), time plots for actual versus forecast are not same. The points in the residual plots are not within the UCL and LCL band. This indicates that the residuals are not random, they are correlated. This is, one indication in turn, these models are inadequate for the given data. Also, it is observed that, the model ARIMA (1,1,0), time plots for actual versus forecast are almost same. Almost all the points in the residual plots are within the UCL and LCL band. This indicates that residual are random. The chi-square statistics computed for groups of lags m=12, 24, 36 and 48 are not significant as indicated by the large p-values. Hence, this model ARIMA (1,1,0) is adequate for the given data. Hence, one can consider the model ARIMA (1,1,0) is the best fit for the given data.*

Keywords: *Auto regressive Integrated Moving Average (ARIMA), Time series, XL Miner, Auto Correlation function (ACF), Partial Auto Correlation function (PACF).*

1. INTRODUCTION

Statistics is probably a much friendlier branch of mathematics because it really can be used every day. Statistical methods have been used in almost every applied field to analyze experimental data. Knowing statistics in everyday life will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected. Even with all the data stored in the largest of data warehouses business decisions still just become more informed guesses.

Data mining it is useful to look at the literal translation of the word to mine in English means to extract. The verb usually refers to mining operations that extract from the earth hidden, precious resources.

Time series is a set of statistical observations arranged in chronological order. A time series may be defined as collection of magnitudes belonging to different time periods, of some variable or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco, or index of industrial production. When quantitative data are arranged in the order of their occurrence the resulting statistical series is called a time series. A time series is a set of observations taken at specified times, usually at equal intervals.

2. METHODOLOGY

The time series models with the add-in package in MS-Excel namely XLMiner. The various methods available in XLMiner for data mining. The various time series techniques, such as ARIMA (Auto Regressive Integrated Moving Average), ACF (Auto Correlation Function), PACF (Partial Auto Correlation Function) and Exponential smoothing are presented in this paper.

Partition data

In the context of DM, the data should be partitioned into training data and validation data. First fit a tentative models and analyses for the training data set and then check the validation with validation data set. Generally, one can partition the data into 60% as training data set and 40% as validation data set.

ARIMA (Auto-Regressive Integrated Moving Average)

The Box –Jenkins methodology refers to the set of procedures for identifying, fitting, and checking ARIMA models with time series data. Forecasts follow directly from the form of the fitted model.

Auto regressive model: AR (p)

The general form of auto regressive,

$$Y_t = \phi_0 + \phi_1 \mathbf{Y}_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t \quad (1)$$

where, Y_t -Response (dependent) variable at time t

$Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ -Response variable at time lags t-1, t-2,…, t-p, respectively.

$\phi_0, \phi_1, \phi_2, ..., \phi_p$ -Coefficients to be estimated

ε_t -Error term at time t

Moving Average Model: MA (q)

The general form moving average,

$$Y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (2)$$

where, Y_t -Response (dependent) variable at time t

μ -Constant mean of the process

$\theta_1, \theta_2, ..., \theta_q$ -Coefficients to be estimated

ε_t -Error term at time t

$\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-q}$ -Errors in previous time periods that are incorporated in the response Y_t

Autoregressive Moving Average Model: ARMA (p, q)

The general form autoregressive moving average,

$$Y_t = \phi_0 + \phi_1 Y_{t-1} \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (3)$$

Autoregressive Integrated Moving Average Model: ARIMA (p, d, q)

The ARIMA (p, d, q) here p indicates the order of the autoregressive part, d indicates the amount of differencing, and q indicates the order of the moving average part. If the ARIMA models. The difference linear operator (Δ), defined by

$$\Delta Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1-B)Y_t \quad (4)$$

The stationary series W_t obtained as the d the difference (Δ^d) of Y_t

$$W_t = \Delta^d Y_t = (1-B)^d Y_t \quad (5)$$

ARIMA (p, d, q) has the general form

$$\phi_p(B)(1-B)^d Y_t = \mu + \theta_q(B)\varepsilon_t \quad or \ \phi_p(B)W_t = \mu + \theta_q(B)\varepsilon_t \quad (6)$$

where, p- Number of AR parameters.

d-Number of differences.

q-Number of MA parameters.

ACF (Auto Correlation Function)

The ACF of an ARMA process $\{X_t\}$ is the function $\rho(\cdot)$ found immediately from the ACVF $\gamma(\cdot)$ as

$$\rho(h) = \frac{\gamma(h)}{\rho(0)} \quad (7)$$

Likewise, for any set of observations $\{x_1,...,x_n\}$, the sample ACF $\hat{\rho}(\cdot)$ is computed as

$$\hat{\rho}(h)=\frac{\hat{\rho}(h)}{\hat{\rho}(0)} \tag{8}$$

The Autocorrelation is the correlation between observations of a time series separated by say, k time units. Suppose there are n time based observations, X_1, X_2, X_3,..., X_n. In ACF technique XLMiner finds correlation between the observations for different lags.

When lag=1, XLMiner creates the following two sets and finds a value of correlation.

X	X_0	X_1	X_2	…	X_{n-1}
Y	X_1	X_2	X_3	…	X_n

Lag=2 gives another such table and one more value of correlation.

X	X_0	X_1	X_2	...	X_{n-2}
Y	X_2	X_3	X_4	…	X_n

If the data is random then the plot should be within the UCL (Upper confidence level) and LCL (Lower confidence level). If it goes beyond UCL or LCL, then one can conclude that some correlation exists in the data.

PACF (Partial Auto Correlation Function)

The partial autocorrelation function (PACF) of an ARMA process $\{X_t\}$ is the function $\alpha(\cdot)$ defined by the equations $\alpha(0)=1$ and $\alpha(h)=\phi_{hh}, \quad h\geq 1,$

where ϕ_{hh} is the last component of

$$\phi_h=\Gamma_h^{-1}\gamma_h \tag{9}$$

$$\Gamma_h=[\gamma(i-j)]_{i,j=1}^{h} \text{ and } \gamma_h=[\gamma(1),\gamma(2),...,\gamma(h)]'$$

For any set of observations $\{x_1,...,x_n\}$ with $x_i\neq x_j$ for some i and j, the sample PACF $\hat{\alpha}(h)$ is given by $\hat{\alpha}(0)=1$ and $\hat{\alpha}(h)=\hat{\phi}_{hh}, \quad h\geq 1$ where $\hat{\phi}_{hh}$ is the last component of

$$\hat{\phi}_h=\hat{\Gamma}_h^{-1}\hat{\gamma}_h. \tag{10}$$

PACF technique is used to compute and plot the partial autocorrelations of a time series. With PACF one can find correlation between some components of the series, eliminating the contribution of other components. PACF technique measures the strength of relationship with other terms being accounted for XLMiner applies these techniques and shows the plots. For this it is good to partition the data into training and validation sets.

3. STATISTICAL ANALYSIS TOOLS

XLMiner is a tool for data analysis in MS-Excel that uses classical and modern computationally intensive techniques. XLMiner is a complete data mining add-in software for MS-Excel. XLMiner is a tool for data analysis in MS-Excel that uses classical and modern computationally intensive techniques. XLMiner is an affordable, easy to use tool for business analysts, consultants and students to learn strengths and weaknesses of data mining methods, prototype large-scale data mining applications and implement medium scale data mining applications.XLMiner is Unique in the content of low cost, comprehensive set of data, mining models and algorithms that includes statistical, machine learning and database methods. Compared to other software, XLMiner provides low learning hurdle. It enables interactive analysis of data, facilitates incorporation of domain knowledge. By empowering of data and post-processing of results using Excel functions, reporting in word, presentations in PowerPoint. Also it supports communication between data miners and end-users, smooth transition from prototyping to custom solution development. Resampling Stats, Inc. distributes XLMiner and one can download it from www.xlminer.net. XLMiner contains the following methods which are used for data mining applications.

- **Partition data:** Standard partition, partition with over sampling
- **Data utilities:** Sample from worksheet, sample from data base, missing data handling, bin continuous data, transform categorical data (create dummies. create category scores, reduce categories).
- **Time series:** Partition data, ARIMA, ACF, PACF, smoothing, (exponential, double exponential, moving average, holt-winter (multiplicative, additive), holt-winter no trend).
- **Prediction:** Multiple linear regression, k-nearest neighbors, regression tree, neural network (multilayer feed forward).
- **Classification:** Discriminant analysis, logistic regression, classification tree, naïve bayes, neural network (multilayer feed forward), k- nearest neighbors
- **Affinity:** Association rules.
- **Data reduction and exploration:** Principal components, k-means clustering, hierarchical clustering.
- **Charts:** Box plot, histogram, and matrix plot.

Time series components in XLMiner

The following methods are available in XLMiner for constructing the time series models.

- Partition data
- ARIMA (Auto-Regressive Integrated Moving Average)
- ACF (Auto Correlation Function)
- PACF(Partial Auto Correlation Function)
- Smoothing, (Exponential, Double Exponential)
- Moving average
- Holt-Winter (Multiplicative, Additive), Holt-Winter no trend.

4. RESULTS AND DISCUSSION

The data set contains 71 observations for the period 1929 to 1999. The results and discussions of the fitted ARIMA models with various parameters are as follows:**ARIMA (1, 1, 0).** First, partition the given data into

training data (60%) and validation data (40%). Construct the exploratory techniques ACF and PACF for both the partitioned data. The ACF and PACF plots for the model ARIMA (1,1,0) are shown in figures 1 and figure 2.

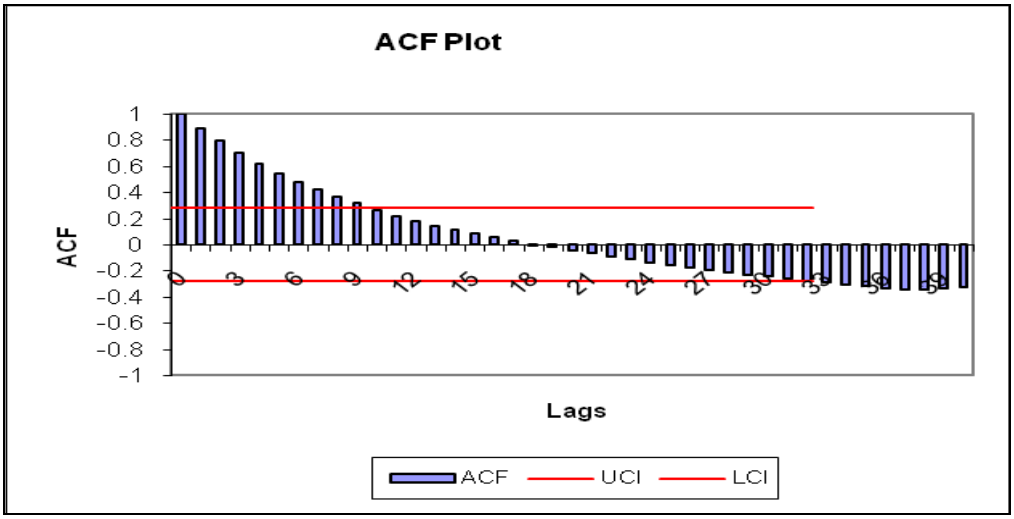

Fig. 1(a) Training data

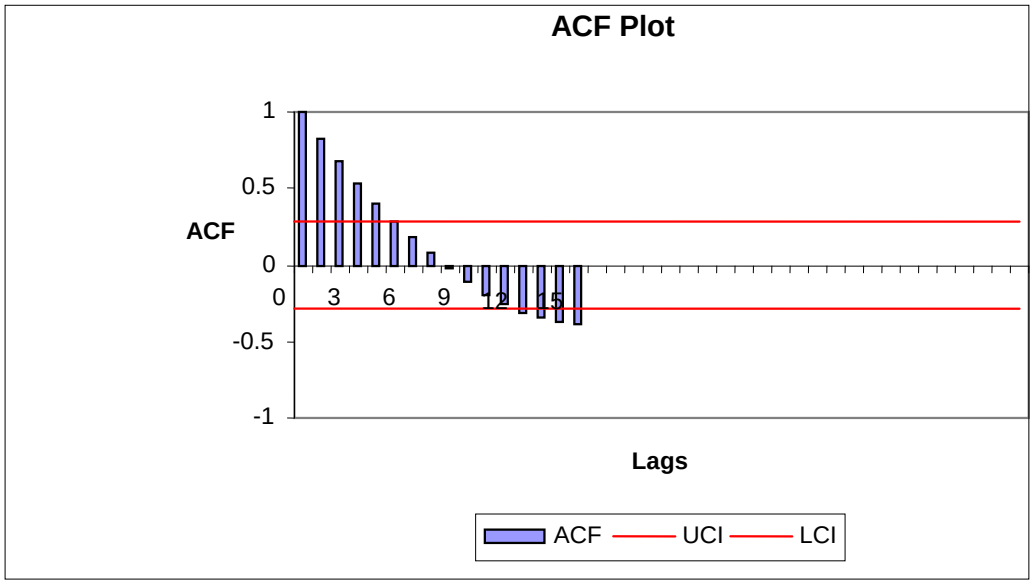

Fig. 1(b) Validation data

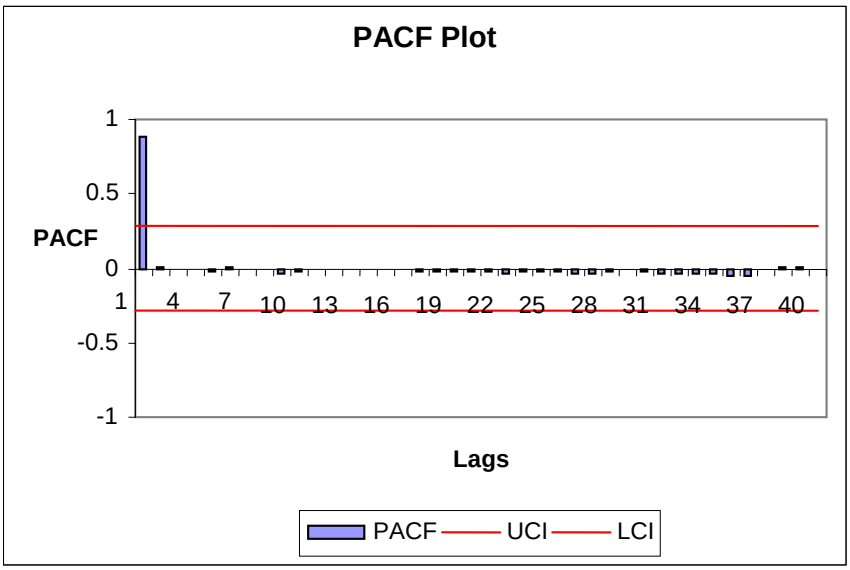

Fig. 2(a) Training data

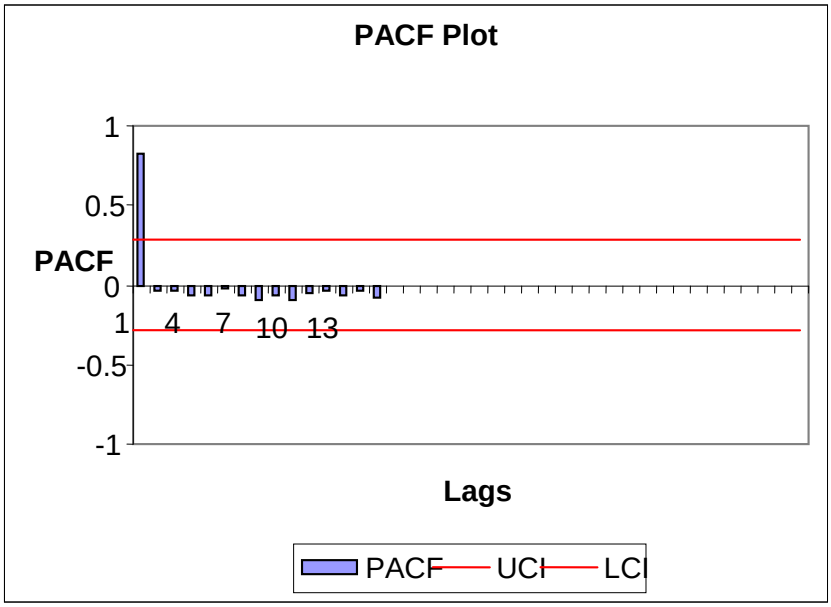

Fig. 2 (b) Validation data

ACF and PACF plots help us to tentatively identify the model and also show if there is a trend and / or seasonality in the data. From the above figures, it is observed that ACF and PACF are similar pattern. Figure 1 shows that, ACF function decreases with the lag increases. This means there is a trend in the data. Since the pattern does not repeat, one can conclude that the data does not show any seasonality and then proceed to fit the model tentatively.

The computed constant and the coefficient term for the model ARIMA (1,1,0) by using XLMiner and is displayed in the following Table 1 and Table 2.

Table. 1 Results based on ARIMA (1, 1, 0)

ARIMA	Coeff	StErr	p-value
Const.	3.825	8.782	0.663
AR1	0.9783	0.0276	0

Table. 2 Comparison of ARIMA Values

Lag	12	24	36	48
p-Value	0.4286	0.4365	0.3454	0.1995
ChiSq	11.1764	23.4181	37.7318	54.9227
df	11	23	35	47

From the above table, the fitted forecasting equation is,

$$Y_t = 3.825 + 0.978\, Y_{t-1} + \varepsilon_t \quad (11)$$

The time plot of the actual versus forecasted values for both training and validation data are shown in the following figure. 3.

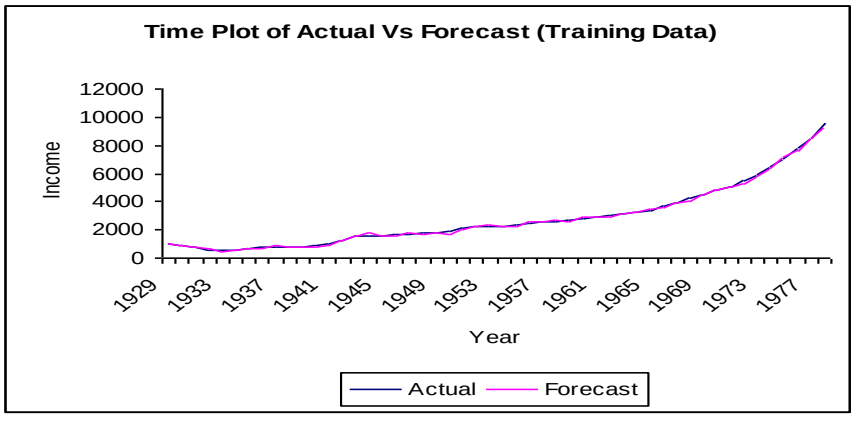

Fig. 3(a) Training data

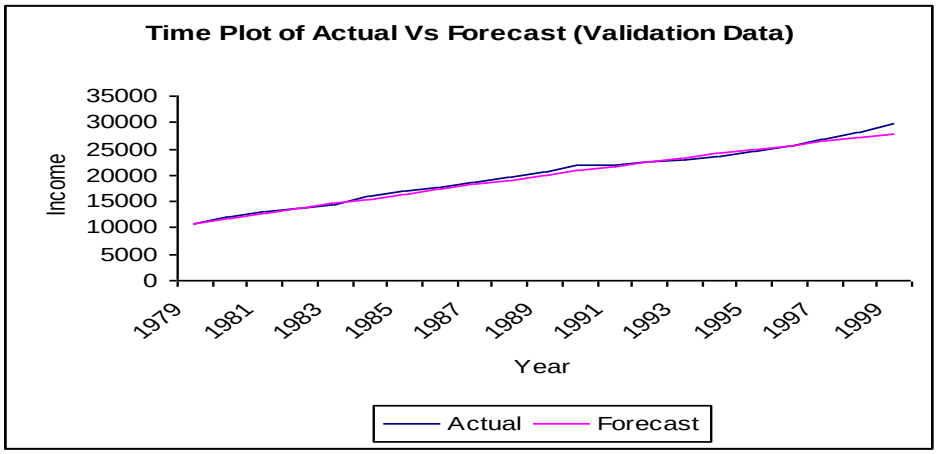

Fig. 3(b) Validation data

From the figure, it is observed that, the actual and predicted values are almost same and in the both cases. XLMiner displays the ACF and PACF plots for residuals which are given in the following figure 4.

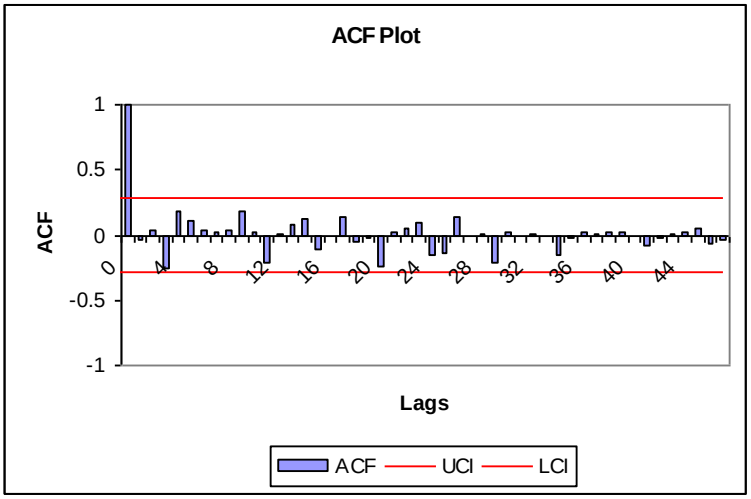

Fig. 4(a) ACF plot for Residuals

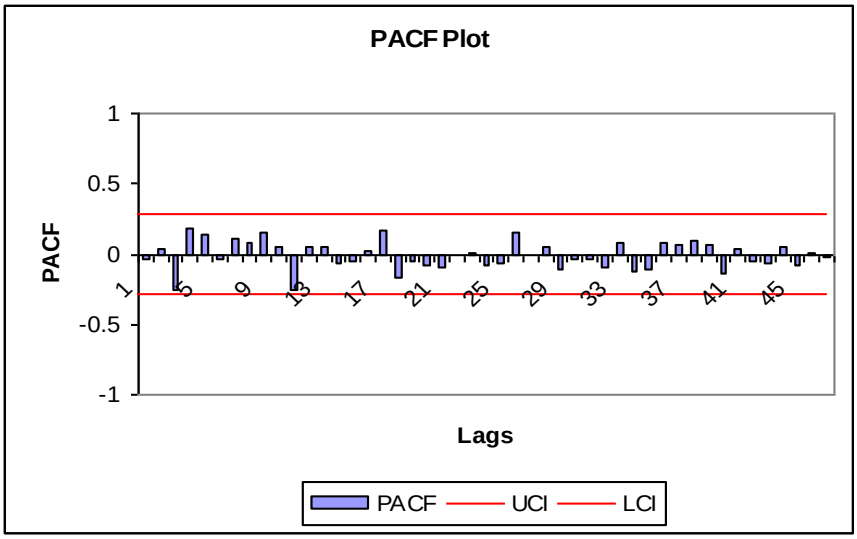

Fig. 4(b) PACF plot for Residuals

The above figure shows that, almost all the points are within the UCL and LCL band. This indicates that the residuals are random, they are not correlated. This is, one indication in turn, that the model ARIMA (1,1,0) is adequate for the given data.

The ARIMA with varying the values of the parameters p,d,q and thus the results obtained using XLMiner are given in the appendix. The following observations are made from the results.

It is observed that, all the models except ARIMA (1,1,0), time plots for actual versus forecast are not same. The points in the residual plots are not within the UCL and LCL band. This indicates that the residuals are not random, they are correlated. This is, one indication in turn, these models are inadequate for the given data.

It is observed that, the model ARIMA (1,1,0), time plots for actual versus forecast are almost same. Almost all the points in the residual plots are within the UCL and LCL band. This indicates that residual are random. The chi-square statistics computed for groups of lags m=12, 24, 36 and 48 are not significant as indicated by the large p-values. Hence, this model ARIMA (1,1,0) is adequate for the given data. Hence, one can consider the model ARIMA (1,1,0) is the best fit for the given data.

6. CONCLUSION

Statistics plays an extremely vital role in several areas of science as an aid to decision making. Statistics has always been about creating methods to analysis data. The increasing availability of data in the current information society has led to the need for valid tools for it modeling and analysis. Data Mining and applied statistical methods are the appropriate tools to extract knowledge from such data.

In this paper, we have discussed the main issues of statistics which are highly relevant to DM and have much to offer to DM. The frequently used statistical methods for data mining are also reviewed. Also, we have discussed elaborately, the data mining with the time series modeling approach through XLMiner. For that we consider two data sets, one for non-seasonal and other for seasonal. We have fitted the ARIMA models with various value of p,d,q parameters. For that, first, partition the given data into training data and validation data. Then, construct explorative techniques ACF and PACF for the both data and identify the tentative model. Fit a

tentative model, and check whether the fitted model is adequate or not through results, ACF and PACF plots for residuals.

The ARIMA with varying the values of the parameters p,d,q and thus the results obtained using XLMiner. It is observed that, for the non-seasonal data all the models except ARIMA (1,1,0), time plots for actual versus forecast are not same. The points in the residual plots are not within the UCL and LCL band. This indicates that the residuals are not random, they are correlated. This is, one indication in turn, these models are inadequate for the given data. Also, it is observed that, the model ARIMA (1,1,0), time plots for actual versus forecast are almost same. Almost all the points in the residual plots are within the UCL and LCL band. This indicates that residual are random. The chi-square statistics computed for groups of lags m=12, 24, 36 and 48 are not significant as indicated by the large p-values. Hence, this model ARIMA (1,1,0) is adequate for the given data. Hence, one can consider the model ARIMA (1,1,0) is the best fit for the given data.

Data Mining has to do with the discovery of useful, valid, unexpected and understandable knowledge from data. These general objectives are obviously shared by other disciplines like statistics, machine learning or pattern recognition. One of the most important distinguishing issues in data mining is size. With the advent of computer technology and information systems, the amount of data available for exploration has increased exponentially. This poses difficult challenges to the standard data analysis disciplines: one has to consider issues like computational efficiency, limited memory resources, interfaces to databases, etc. All these issues turn data mining into a highly interdisciplinary subject involving tasks not only of typical data analysts but also of people working with databases, data visualization on high dimensions, etc. The criteria for Data Mining and modern data analysis software are (i) Easy-to-use when handling data and (ii) Comprehensive Range of procedures such as Regression, multivariate procedures, neural nets, and classification trees.

The software which are designed for data mining purpose are, SAS Enterprise Miner, SPSS Clementine, IBM Intelligent Miner and etc., These software are powerful, comprehensive, easy-to-use; but it needs substantial learning effort and expensive.

XLMiner is unique in the content of low cost, comprehensive set of data, mining models and algorithms that includes statistical, machine learning and database methods. Compared to other software, XLMiner provides low learning hurdle. It enables interactive analysis of data, facilitates incorporation of domain knowledge. The main limitation of XLMiner is size; holding of large data an Excel spreadsheet is not possible. If Excel is used as a view-port into a database such as MS-Access, My SQL server and Oracle, these limits do not apply.

Data mining is an emerging discipline that is used to extract information from large databases. Now-a-days, data collected and stored at enormous speeds (GB/Hr). Lots of data are being collected and warehoused (for example, web data, and credit card transactions) and pressure to provide better, customized services for an edge. In this context, traditional techniques are infeasible due to enormity of data, high dimensionality of data and heterogeneous of data. By considering the issues of statistics related to Data Mining, there is an increasing need for valid tools instead of traditional statistical procedures, which can deal with ever larger data sets.

7. REFERENCES

1. Benjamini and Hochberg,, "Controlling the false discover rate : A practical and powerful approach to multiple testing", *.J.R.Statist.Soc.*B, 57,289-300.
2. David.Hand J, "Data Mining: Statistics and More", *American Statistical Association,* 52, 112-118, 1998.
3. David Hand J, "Statistics and Data Mining: Intersecting Disciplines", *SIDKDD Explorations,* 1(1), 16-19, 1999.
4. David Hand J & et al. Principles of Data Mining, *Eastern Economy Edition,*2001.
5. Elhance D.N, Veena Elhance B.M & Aggarwal, Fundamentals of Statistics.
6. George E.P.Box, Gwilym M. Jenkins & et al., "Time Series Analysis Forecasting and Control", *Low Price Edition*, 2004.
7. Han.J & M.Kambar, Data Mining Concepts and Techniques, *Morgan Kanufann Publishers,* 2001
8. Jiawei Han & Micheline Kambar , Data Mining Concepts and Techniques, *Elsevier, 2006*
9. John E.Hanke & Dean W.Wichern, Business Forecasting, *8th Edition,* 2009
10. Leshno M, & et al, "Multilayer feed forward networks with a non polynomial activation function can approximate any function", *Neural Networks,* 6, 861-867, 1993
11. McCullagh & Nelder, Generalized Linear Model, Chapman and Hall, 1991
12. Osmar R.Zaiane, Introduction to Data Mining, *CMPUT690 principles of knowledge Discovery in Databases,* 1999
13. Paolo Giudici, "Applied data mining statistical methods for business and industry", 2003
14. Pang-Ning Tan, Michael Steinbach & et al, "Introduction to Data Mining", 2009
15. Peter J.Brockwell & Richard A.Dawis, "Introduction to Time Series and Forecasting", *Springer International Edition,* 2002
16. Rangsan Nochai & Titida Nochai, "ARIMA Model For Forecasting Oil Palm Price", Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications, University Sains Malayasia, Penang, 2006
17. Trevor Hastie & et al., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", *Springer-Verlag,* 2008
18. Yoav Benjamini, "Statistical Methods for Data Mining", 555-587, 2001
19. www.XLMiner.com