

MODELING AND FORECASTING TIME SERIES: USING THE ARIMA MODEL TO PREDICT RAINFALL

^{1*} Biswaranjan Tripathy, ² Biswa Ranjan
^{1*} Professor, Dept. Of Civil Engineering, NIT BBSR,
Asst. Professor Dept. of Civil Engineering, SEARC, BBSR
^{1*} tripathy.b8@gmail.com, biswaranjan@gmail.com

Abstract: In this paper, a Time Series Modeler (TSM) for rainfall forecasting in an Indian coastal region is presented. A five-year dataset (2009–2013) with primary attributes including temperature, dew point, wind speed, maximum temperature, minimum temperature, visibility, and rainfall was used to create this model. The Statistical Package for Social Studies (SPSS) TSM has been used as an innovative approach for training and testing this dataset. A reliable model for rainfall prediction is thus feasible because the performance criteria for this model's evaluation are based on the significant values of the statistical performance measures, namely mean absolute deviation (MAD), mean squared error (MSE), mean absolute percent error (MAPE), and root mean square error (RMSE). The prediction accuracy range of the outcomes produced by this model is substantially within acceptable bounds at 80%. This model is based on the SPSS 20.0 TSM auto regressive integrated moving average (ARIMA) model.

Keywords: auto regressive integrated moving average; ARIMA; Statistical Package for Social Studies; SPSS; Time Series Modeler; TSM; time series data; modelling; statistical measures; weather forecast; rainfall prediction; forecast performance measures.

Introduction

Knowing what might happen to a system in the upcoming time periods is a phenomena known as forecasting. Temporal forecasting, also known as time series prediction (Imdadullah, 2014), anticipates future values for a sequence of data with values $x_t - n, \dots, x_t - 2, x_t - 1, x_t$. The objective is to monitor or model the current data series in order to properly predict future unknown data values. The attributes needed to predict rainfall are so complicated because weather data is continuous, data-intensive, and dynamic (Geetha and Nasira, 2014a), therefore even short-term predictions are subject to error. Forecasting rainfall is extremely difficult because of these distinctive characteristics. The prediction of rainfall is done using a variety of methodologies, including data mining, fuzzy logic, evolutionary algorithms, and statistical methods (Banu and Tripathy, 2016). (Sharma et al., 2014). The focus of this paper is on statistical TSM approaches utilising IBM SPSS Statistics 20.0. (Schiopu et al., 2009). By modelling based on the correlations in the weather forecasting data, the auto regressive integrated moving average (ARIMA) model (Li et al., 2013) is a purely statistical method for analysing and developing a forecasting model that best represents a time series (Babu et al., 2015). Numerous benefits of the ARIMA model were discovered in the empirical research, which supports the ARIMA for forecasting short-term time series. The ARIMA method (Zakaria et al., 2012) generalises the forecast using only the prior past data of a rainfall time series, taking advantage of its strictly statistical approach. Consequently, the ARIMA approach can improve forecasts

accuracy while keeping the number of parameters to a minimum. Thus, the objective of this paper is to design a model as a disaster prediction system (Devi et al., 2013; Kusumastuti, 2014).

1 ARIMA model

The time series is represented in the real time world, as follows

$$X(t-a) \dots X(t-2), X(t-1), X(t)$$

For time series prediction, there are many numerical methods, but we analyse and predict based on the previous historical data. For the past N samples, it is can be represented as

$$\hat{Y}(n+1) = \sum ai.x(n-i)$$

where the prediction coefficient is $ai, i = 0, 1, 2 \dots \dots N - 1$.

ARIMA model is popularised by Box and Jenkins. It is a combination of three mathematical models namely auto-regressive, integrated, moving-average (ARIMA) models of time series data. Time series analysis is a set of observations observed at a particular time period. An ARIMA (p, d , and q) model can account for temporal dependence in several ways, where p is the order of the autoregressive part, d is the order of the differencing and q is the order of the moving-average process.

- First, the time series considers being stationary, by taking d differences. If $d = 0$, i.e., no differencing is done, the models are usually referred to as ARMA (p, q) and the observations are modelled directly. If $d = 1$, the differences between consecutive observations are modelled.

- Second, term is autoregressive, which is capable of wide variety of time series forecasting by adjusting the regression coefficients. Since the independent variables are time-lagged values for the dependent variable, the assumption of uncorrelated error is easily violated. The equation is given by,

$$X_t = a + \sum \varphi_i X_{t-i} + \varepsilon_t$$

where a is the constant, φ_i is the parameter of the model, x_t is the value that observed at t and ε represents random error and i varies from 1 to p .

- Third, q is the moving-average term; the basic idea of Moving-Average model is finding the mean for a specified set of values and then using it to forecast the next period and correcting for any mistakes made in the last few forecasts. The equation is:

$$X_t = \varepsilon_t + \sum \theta_i \varepsilon_{t-i}$$

where θ_i is the parameter of the model, ε_t is the error term and i varies from 1 to q .

- Combining these three models we get ARIMA (p, d, q) model, it uses combinations of past values and past forecasting errors and offer a potential for fitting models that could not be adequately fitted by using an AR or an MA model alone. Furthermore, the addition of the differencing eliminates most non-stationarity in the series. So, the general form of the ARIMA models is given by

$$Y_t = a_0 + \sum \varphi_i \cdot Y_{t-i} + \sum \theta_i \cdot \varepsilon_{t-j}$$

where Y_t , a stationary is a stochastic process, a_0 is the constant, ε_t is the error or whitenoise disturbance term, φ_i means auto-regression coefficient and θ_i is the moving average coefficient, where $i \in 1$ to p and $j \in 1$ to q .

The flexible nature of the ARIMA model (for both seasonal and non-seasonal models), motivated us that our weather dataset, which is highly dynamic, chaotic and multi dimensional aptly fits for ARIMA (Yadav and Balakrishnan, 2014), which provides us a solid foundation, as there is always uncertainty and gamble in weather prediction (Geetha and Nasira, 2014b). An ARIMA model (Rahman et al., 2013) can be viewed as a 'filter' as it tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts.

Time-series model

Any data collected over a period of time is called time series data. There are many benefits of time series data. A time-series (Gupta et al., 2013) is a collection of observations made sequentially through time. Thus, a time series is a set of observations obtained by measuring a single variable or multiple variables regularly over a period of time. One of the most important objectives of time series analysis (Nury et al., 2013) is to forecast future values of the series called as time series forecasting Adela (2013).

- to analyse the behaviour of the past data
- to forecast the future series
- to compare and contrast
- to evaluate the trend in the series

as a control standard for a parameter. The two basic models for time domain are

- 1 ARIMA model
- 2 Regression model (Geetha and Nasira, 2014c).

As IBM SPSS 20.0 supports time series data as well as ARIMA, it is considered ideal for weather prediction (SriPriya and Geetha, 2015) particularly rainfall. Because of the features of SPSS like wizards, multiple tab options with all the mandatory and optional categories, output panes, zoom and plot windows, graphical and descriptive representations made us to stick on to SPSS. Designing the model, efficiency and accuracy of SPSS are the main significant factors for selecting this tool. The other tools in the market are

- SAS
- R
- NCSS
- Orange.

Good forecasts and modelling (Majumdar, 2010) are vital in many areas of scientific, industrial, commercial, marketing, financial (Radhwan et al., 2015), sales, medical, share trading and any other economic activities. Our weather (rainfall) dataset is an ideal example of time series data (Filzah et al., 2013). Weather data are available from authentic organisations and resources where, observations of hourly, weekly, monthly, quarterly, half yearly, yearly, century-wise are available with many attributes.

3 Literature review

Weather forecasting (Geetha and Nasira, 2014c) is a fascinating phenomenon of Meteorology and has been one of the most challenging problems around the world because of its day today usage in common man's regular activities to a satellite launch expert or to aviation personnel. Weather forecasting is a widely played popular magic cube for scientific research and development, especially for prediction of rainfall.

Few scientific research works related to the weather forecasting are highlighted. Fuzzy logic is widely used in the atmospheric variables, data analysis and prediction. Schiopu et al. (2009) tried factor analysis and linear regression and concluded that factor analysis reduces large number of variables into less factors using SPSS statistical methods.

Singh et al. (2011) proposed the use of the time series based temperature prediction model using integrated back propagation/genetic algorithm techniques. Gupta et al. (2013) tried time series analysis of forecasting Indian rainfall and concludes that back propagation neural network was acceptably accurate and can be used for predicting the rainfall. Sasu (2013) made a quantitative comparison of models for univariate time series forecasting using ARIMA model and IBM SPSS.

Li et al. (2013) implemented Hadoop-based ARIMA Algorithm which has the ability of mass storage of meteorological data, efficient query and analysis, weather forecasting and other functions. Rahman et al. (2013) made a comparative study on ANFIS and ARIMA model for weather forecasting in Dhaka and concluded that ARIMA is efficient for temperature forecasting. Geetha and Nasira (2014b) successfully implemented artificial neural networks (ANNs) for rainfall prediction using RapidMiner tool to produce an accuracy percentage of 82%. They have supplemented the paper with the steps to implement, input and output screen shots and had plotted a graph by comparing the actual and the predicted values. Patel et al. (2014) implemented and concluded that as error is very less, ARIMA model is best to predict rain attenuation for Ku-band satellite for 12 GHz frequency.

Babu et al. (2015) stated that ARIMA is most effective method for weather forecasting than ANFIS, but ANFIS consumes less time for processing than ARIMA. SriPriya and Geetha (2015) in their paper had made a pilot study to predict the tropical cyclones of India, using Chi-Square Automatic Interaction Detector (CHAID) decision tree. They have used nearly 14 storm attributes, and trained using three years dataset to predict for the next consecutive year. They are successful in predicting upto 90% accuracy. SriPriya and Geetha (2015) in their paper, had made a significant contribution

by predicting Storms using the Data Mining tool R, using K-NN algorithm. The challenge is the proper selection of the machine learning technique to get accurate prediction using only the three types of input weather variables: estimated central pressure, maximum sustained surface wind and pressure drop.

4 Case study: rainfall data analysis of Trivandrum

Trivandrum is situated in the south west coast of Kerala. The climate of Trivandrum is hot tropical. The Trivandrum District gets rainfall from both the south-west Monsoon and the north-east Monsoon. It is situated between north latitudes 8°17' and 8°54' and east longitudes 76°41' and 77°17'. In this paper, we have collected the weather dataset from the site <http://ftp.ncdc.noaa.gov/pub/data/gsod/2009-2015/>. The station code 433710 refers to the location Trivandrum.

Figure 1 Rainfall data of Trivandrum (see online version for colours)

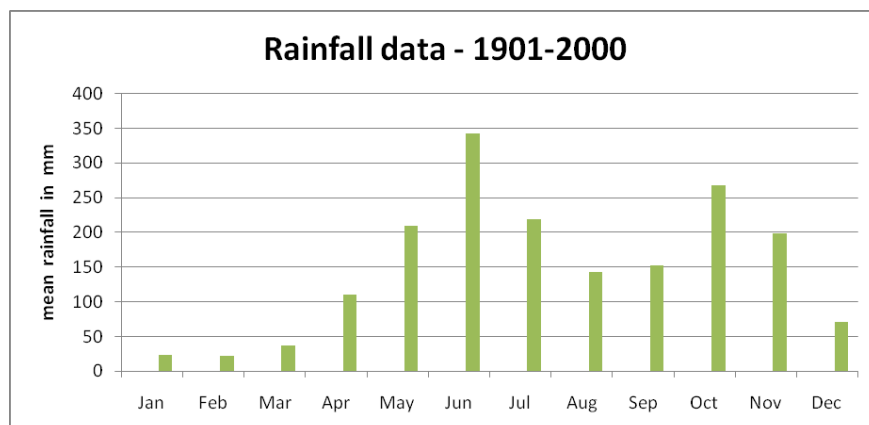


Figure 1 depicts a graphical representation of rainfall data (1901–2000) of Trivandrum. Courtesy: <http://www.imd.gov.in/doc/climateimp.pdf>. The south-west monsoon sets in by June and lasts by the month of September whereas the north-east monsoon starts in October and fades by November. It is the first city along the path of the south-west monsoon and gets its first showers in early June.

5 Implementation of TSM using ARIMA model

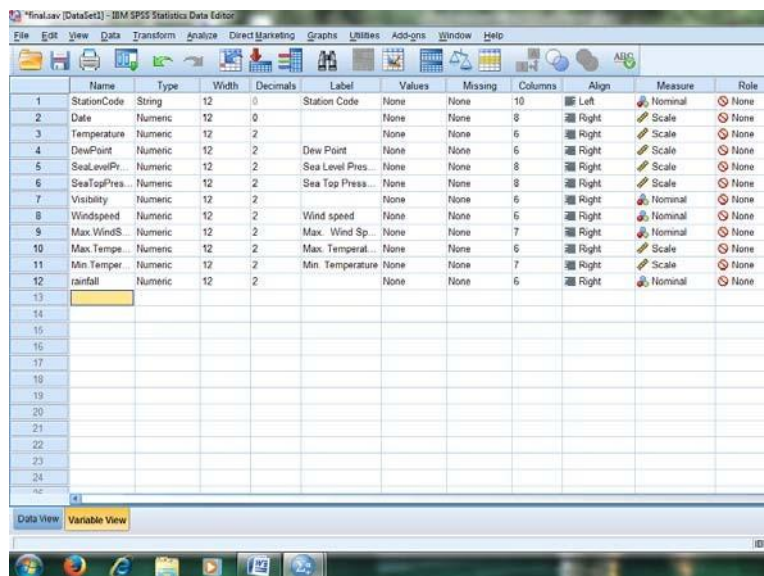
Building a model to forecast

The Forecasting module of TSM provides two procedures for accomplishing the task of creating models and producing forecasts. The Expert Modeler of TSM automatically determines the best mode for time series weather data. Table 1 depicts rainfall dataset along with its description and Figure 2 in SPSS.

Table 1 Rainfall dataset description

no.	tribute	Type	Description
1	IN	String	Station code
2	ATE	Numeric	Year, month, day
3	EMP	Numeric	Mean Temperature in F
4	EWP	Numeric	Mean dew point in F
5	_P	Numeric	Mean sea level pressure in mb
6	FP	Numeric	Mean station pressure in mb
7	ISIB	Numeric	Mean visibility in miles
8	'DSP	Numeric	Mean wind speed in knots
9	XSPD	Numeric	Maximum sustained wind speed in knots
10	AX	Numeric	Maximum temperature in F
11	IN	Numeric	Minimum temperature in F
12	AINFALL	Numeric	Total precipitation in inches

Figure 2 Screen shot of weather dataset (see online version for colours)



Implementation procedure of ARIMA model

We have to determine whether our rainfall dataset (2009–2013) exhibits seasonal variations. Only based on that, we can conclude, whether the dataset is fit for TSM. This is done by selecting through the choices from the menu bar, Analyse → Forecasting → Sequence charts.

Figure 3 Screen shot of sequence chart

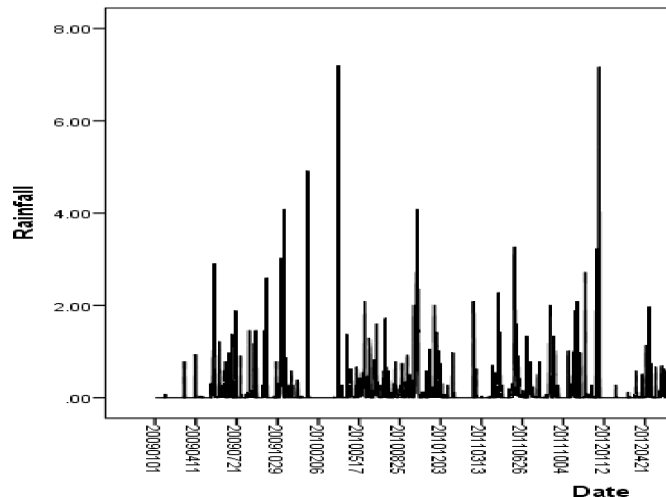
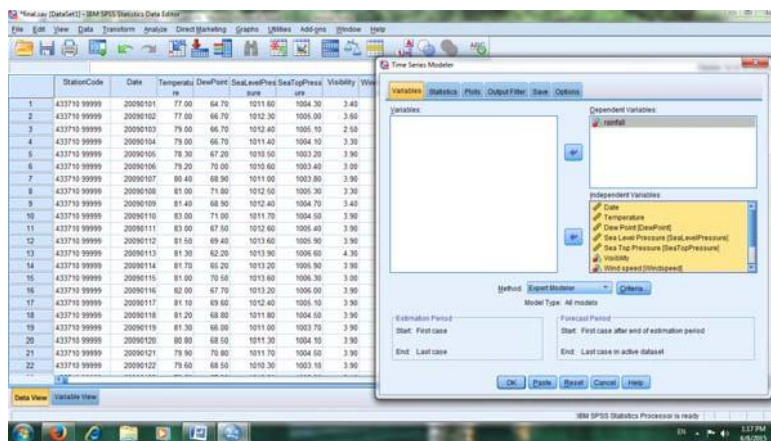


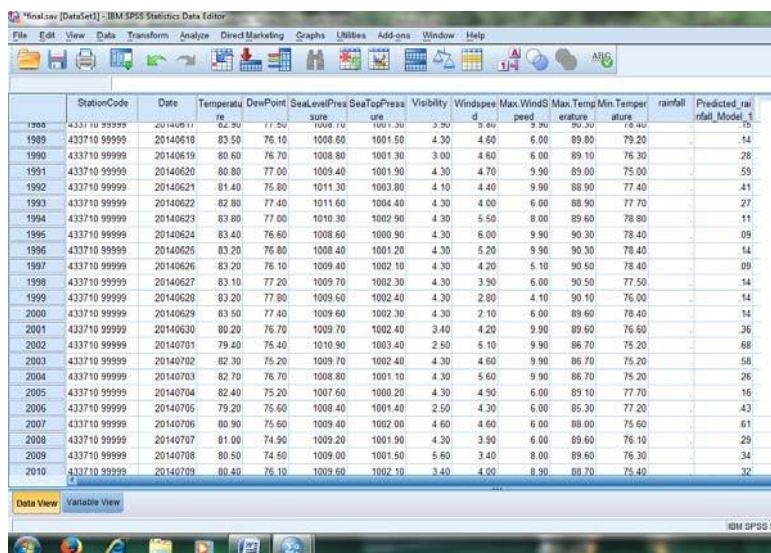
Figure 3 stands as a strong evidence to create the model, as there is no seasonal periodicity. As the dataset is ideal for TSM, we then preprocessed the data by replacing the missing values with the mean values, so that the dataset is normalised. To create the model, as in Figure 4, i.e., to use the Expert Modeler, Analyse → Forecasting → Create Models.

Figure 4 Time Series Modeler window (see online version for colours)



The model is trained by using the five years dataset from the year 2009–2013 with all the 12 weather attributes. And the model is tested with 2014 data excluding the attribute rainfall.

Figure 5 Screen shot with predicted rainfall model_1 for 2014 dataset (see online versionfor colours)



Thus, we have created our model and predicted rainfall for the year 2014, as depicted in Figure 5. Also, SPSS 20.0 offers another feature named ‘Apply Model’ which extends the forecasts without rebuilding the model again. Analyse → Forecasting → Apply model.

6 Model validation

The statistical measures of the results are discussed to evaluate the performance of our ARIMA model, which is based on forecast errors. Forecast error is calculated by finding the difference between the actual and the predicted value at a given time period, as shown in the formula,

$$Error\ t = (Actual\ t - Forecast\ t)$$

where *t* is at any given time period.

The commonly used forecast performance measures for summarising historical errors are

- 1 mean absolute deviation (MAD)
- 2 mean squared error (MSE)
- 3 mean absolute percent error (MAPE)
- 4 root mean squared error (RMSE).

These measures enable us to compare the accuracy and among other alternative forecasting methods by determining the one which yields the lowest MAD, MSE, RMSE or MAPE for a given set of data.

Table 2 Model summary

<i>t</i> statistic	Mean	Minimum	Maximum
Stationary R-squared	.205	.205	.205
-squared	.205	.205	.205
MSE	.464	.464	.464
MAPE	340.494	340.494	340.494
MaxAPE	10,471.417	10,471.417	10,471.417
MAE	.217	.217	.217
MaxAE	6.653	6.653	6.653
Normalised BIC	-1.496	-1.496	-1.496

The model fit table as tabulated in Table 2 provides fit statistics calculated across all of the models. It provides a concise summary of how well the models, with re-estimated parameters, fit the data. For each statistic, Table 2 provides the mean, standard error (SE), minimum, and maximum value across all models. While a number of statistics are reported, we will focus on two: MAPE and maximum absolute percentage error (MaxAPE). In statistics, BIC stands for Bayesian information criterion, the model with the lowest BIC is preferred. Based on the significant values we can arrive at a conclusion of building a good model.

Table 3 Model statistics

Model
Number of

Model fit statistics

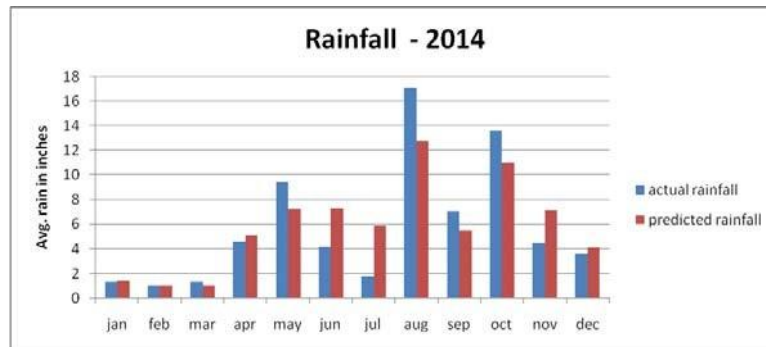
Ljung-Box Q(18)

Number of

Statistics	DF	Sig.outliers				
rainfall-Model_1	5	.205	19.969	16	.222	0

The model statistics table as in Table 3 provides summary information and goodness-of-fit statistics for each estimated model. Results for each model are labelled with the model identifier provided in the model description table. The model contains five predictors out of the 11 candidate predictors that were originally specified. So it appears that the Expert Modeler has identified five independent variables that may prove useful for forecasting. DF means degrees of freedom. A significance value less than 0.05 implies that there is structure in the observed series which is not accounted for by the model. The value of 0.222 shown here is not significant, so we can be confident that the model is correctly specified. Outliers are extreme values far away from the rest of the data, usually they are excluded and here it is nil.

Figure 6 Comparison chart of actual and predicted rainfall (see online version for colours)



7 Conclusions

This paper has demonstrated the prediction of rainfall using ARIMA model of SPSS Time Series Modeler. Our work is promising and encouraging based on the significant values of the statistical indicators RMSE = .464, stationary $R^2 = .205$, MAE = .217 and MAPE = 340.494. Also, by comparing the predicted with the observed values for the years 2014, it is found that the forecast accuracy lies nearly and above 80%. The limitation of ARIMA is, it is strictly statistically based, consumes time, and it is referred as 'backward looking'. But, it yields more accuracy percentage, widely used and has a history of wide acceptance. Thus, the significant value of the statistical indicators challenges us to reach out for higher accuracy.

In future, with the potential of SPSS, predictive analytics can play a vital role in disaster management system, as this work can be extended for predicting floods, land slides, cyclones, earth quakes, tsunamis. Thus, this work has a wider scope as a natural disaster and mitigation system in future.

References

- Babu, R.N., Babu, B.A.C., Reddy, D.P. and Gowtham, M. (2015) 'Comparison of ANFIS and ARIMA model for weather forecasting', *Indian Journal of Science and Technology*, January, Vol. 8, No. S2, pp.70–73.
- Banu, S.K. and Tripathy, B.K. (2016) 'Rough set based similarity measures for data analytics in spatial epidemiology', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 3, No. 1, p.123, DOI: 10.4018/IJRSDA.2016010107.
- Devi, R.B., Rao, N.K., Setty, P.S. and Rao, N.M. (2013) 'Disaster prediction system using IBM SPSS data mining tool', *International Journal of Engineering Trends and Technology (IJETT)*, August, Vol. 4, No. 8, p.33523357.
- Filzah, N., Radzuan, M., Othman, Z. and Bakar, A.A. (2013) 'Uncertain time series in weather prediction', *ICEEI 2013, ScienceDirect, Procedia Technology*, Vol. 11, pp.557–564.
- Geetha, A. and Nasira, G.M. (2014a) 'Rainfall prediction using logistic regression technique', *CiiT International Journal of Artificial Intelligence Systems and Machine Learning*, Vol. 6, No. 7, pp.246–250, ISSN 0974-9667.
- Geetha, A. and Nasira, G.M. (2014b) 'Artificial neural networks' application in weather forecasting – using RapidMiner', *International Journal of Computational Intelligence and Informatics*, Vol. 4, No. 2, pp.177–182.
- Geetha, A. and Nasira, G.M. (2014c) 'Data mining for meteorological applications: decision trees for modeling rainfall prediction', *IEEE Explore*, Print ISBN: 978-1-4799-3974-9, DOI: 10.1109/ICCIC.2014.7238481.
- Gupta, A., Gautam, A., Jain, C., Prasad, H. and Verma, N. (2013) 'Time series analysis offorecasting Indian rainfall', *IJIES*, May, Vol. 1, No. 6, pp.42–45, ISSN: 2319–9598.
- Imdadullah, M. (2014) 'Time series analysis', *Basic Statistics and Data Analysis*, itfeature.com. <http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecasting>.

- Kusumastuti, D. (2014) 'Identifying competencies that predict effectiveness of disaster managers at local government', *International Journal of Society Systems Science*, Vol. 6, No. 2, pp.159–176.
- Li, L., Ma, Z., Liu, L. and Fan, Y. (2013) 'Hadoop-based ARIMA algorithm and its application in weather forecast', *International Journal of Database Theory and Application*, Vol. 6, No. 5, pp.119–132 [online] <http://dx.doi.org/10.14257/ijda.2013.6.5.11> (accessed 16 May 2015).
- Majumdar, P.K. (2010) 'Modelling of coastal hydrogeology of Krishna delta in India', *International Journal of Society Systems Science*, Vol. 2, No. 4, pp.351–374.
- Nury, A.H., Koch, M. and Alam, M.J.B. (2013) Time series analysis and forecasting of temperatures in the Sylhet Division of Bangladesh', *Proceedings of 4th International Conference on Environmental Aspects of Bangladesh*, Fukoka, Japan, August 2013.
- Patel, D.P., Patel, M.M. and Patel, D.R. (2014) 'Implementation of ARIMA model to predict rain attenuation for KU-band 12 Ghz frequency', *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, Vol. 9, No. 1, pp.83–87, ISSN: 2278-8735.
- Radhwan, A., Kamel, M., Dahab, M.Y. and Hassanien, A.E. (2015) 'Forecasting exchange rates: a chaos-based regression approach', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 2, No. 1, p.57, DOI: 10.4018/ijrda.2015010103.
- Rahman, M., Saiful Islam, A.H.M., Nadvi, S.Y.M. and Rahman, R.M. (2013) 'Comparative study of ANFIS and ARIMA model for weather forecasting in Dhaka', 978-1-4799-0400-6/13/\$31.00 © IEEE, *Proceedings of 2nd ICIEV*, 17th–18th May 2013.
- Sasu, A. (2013) 'A quantitative comparison of models for univariate time series forecasting', *Bulletin of the Transilvania University of Brasov*, Vol. 6(55), No. 2, pp.117–124, Series III: Mathematics, Informatics, Physics.
- Schiopu, D., Petre, E.G. and Negoiană, C. (2009) 'Weather forecast using SPSS Statistical Methods', *BULETINUL Universităţii Petrol – Gaze din Ploiesti*, Vol. 61 No. 1, pp.97–100, Seria Matematică – Informatică – Fizică.
- Sharma, M., Mathew, L. and Chatterji, S. (2014) 'Weather forecasting using soft computing and statistical techniques', *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, July, Vol. 3, No. 7, pp.11285–11290.
- Singh, S., Bhambri, P. and Gill, J. (2011) 'Time series based temperature prediction using backpropagation with genetic algorithm technique', *International Journal of Computer Science*, Vol. 8, No. 5, p.3, ISSN: 1694-0814, 2011.
- SriPriya, P.V. and Geetha, A. (2015) 'Cyclone storm prediction using KNN algorithm', *Indian Journal of Engineering*, September, Vol. 12, No. 30, pp.350–354, ISSN 2319-7757, ISSN 2319-7765, Discovery Publication.
- Yadav, R. and Balakrishnan, M. (2014) 'Comparative evaluation of ARIMA and ANFIS formodeling of wireless network traffic time series', *EURASIP Journal on Wireless Communications and Networking*, No. 15, pp.1–8.
- Zakaria, S., Al-Ansari, N., Knutsson, S. and Al-Badrany, T. (2012) 'ARIMA models for weekly rainfall in the semi-arid Sinjar District at Iraq', *Journal of Earth Sciences and Geotechnical Engineering*, Vol. 2, No. 3, pp.25–55, ISSN: 1792-9040 Print, 1792-9660, [online] Science Press Ltd.