OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Time Series Data Analysis for Forecasting – A Literature Review

Neelam Mishra[1], Er. Abhinav Jain[2]

[1](Computer Science, C.S.E., GBTU, India)
[2] (Department of Computer Science, C.S.E., GBTU, India

Abstract: *In today's world there is ample opportunity to clout the numerous sources of time series data available for decision making. This time ordered data can be used to improve decision making if the data is converted to information and then into knowledge which is called knowledge discovery. Data Mining (DM) methods are being increasingly used in prediction with time series data, in addition to traditional statistical approaches. This paper presents a literature review of the use of DM and statistical approaches with time series data, focusing on weather prediction. This is an area that has been attracting a great deal of attention from researchers in the field.*

Keywords: *Data Mining; Time Series Data Analysis; Knowledge Discovery; Weather Prediction*

I. Introduction

Data Mining (DM) is a challenging field for research and has some practical successful application in several different areas. DM methods are being increasingly used in prediction with time series data, in addition to traditional statistical approaches [1-3].

DM can be presented as one of the phases of the Knowledge Discovery in Databases (KDD) process [4-6], and is identified as “the means by which the patterns are extracted from data” [7]. Nowadays, it can be said that the two terms, DM and KDD, are indistinctly used.

A time series is a collection of data recorded over a period of time—weekly, monthly, quarterly, or yearly. An analysis of history—a time series—can be used by management to make current decisions and plans based on long-term forecasting. One usually assumes that past patterns will continue into the future. Long-term forecasts extend more than 1 year into the future; 5-, 10-, 15-, and 20-year projections are common. Long-range predictions are essential to allow sufficient time for various departments to develop plans for future development.

There are several application domains of DM with time series data, being that one important application domain is weather prediction. This will be the focus of this paper. Rainfall prediction is a sensitive issue and can be considered as an open research issue [9,10]. Intelligent forecasting models have achieved better results than traditional statistical methods. Although intelligent forecasting methods are better, we can still improve the results in terms of accuracy in addition to other factors.

The main contribution of this paper is to provide brief literature survey of the use of statistical methods for time series weather predictions, along with the latest trends in the use of DM for time series forecasting

The paper is organized as follows: a review of literature survey on statistical techniques for time series forecasting is presented in section 2 and a literature review on the use of DM with time series data is presented in Section 3. The paper closes in Section 4, with conclusion and future research directions.

II. Statistical Techniques For Time Series Forecasting

Various organizations/ employees in India and abroad have done modeling using supported time series data exploitation. The various methodologies viz. statistic decomposition models, Exponential smoothing models, ARIMA models and their variations like seasonal ARIMA models, vector ARIMA models using variable time series, ARMAX models i.e. ARIMA with instructive variables etc has been used.. Many studies have taken place within the analysis of pattern and distribution of rainfall in numerous regions of the globe. Totally different time series methods with different objectives are used to investigate rain information in numerous literatures.

Stringer (1972) reported that a minimum of thirty five quasi-periods with over one year long are discovered in records of pressure, temperature, precipitation, and extreme climatic conditions over the earth the earth of the world surface. a very common quasi-periodic oscillation is the quasi-biennial oscillation (QBO), during which the environmental condition events recur each two to two.5 years.

Winstanley (1973a, b) reported that monsoon rains from Africa to India decreased by over five hundredth from 1957 to 1970 and expected that the long run monsoon seasonal rain, averaged over five to ten years is probably going to decrease to a minimum around 2030.

Laban (1986) uses time series supported ARIMA and Spectral Analysis of areal annual rain of two same regions in East Africa and counseled ARMA(3,1) because the best appropriate region indice of relative wetness/dryness and dominant quasi-periodic fluctuation around 2.2-2.8 years,3-3.7 years,5-6 years and 10-13 years.

Harvey et al., (1987) investigated how patterns of rainfall correlate with general climatic conditions and frequency of the cycles of rain. They used rain information from Brazil for a selected region which frequently suffers from drought to assess the alternate behaviour of rain. They used a model that permits alternate parts to be modeled explicitly. They found that cyclical components are random instead of deterministic, and also the gains achieved from forecast by taking account of the cyclic element are tiny within the case of Brazil.

Kuo and Sun, (1993) used an intervention model for average10 days stream flow forecast and synthesis that was investigated by to influence the extraordinary phenomena caused by typhoons and alternative serious abnormalities of the weather of the Tanshui river basin in Taiwan.

Chiew et al, (1993) conducted a comparison of six rainfall-runoff modeling approaches to simulate daily, monthly and annual flows in eight unregulated catchments. They concluded that time-series approach will offer adequate estimates of monthly and annual yields within the water resources of the catchments.

Langu, (1993) used statistic analysis to observe changes in rainfall and runoff patterns to go looking for important changes within the parts of variety of rainfall statistic.

Box and Jenkins (1994), in early 1970's, pioneered in evolving methodologies for statistic modeling within the univariate case often referred to as Univariate Box-Jenkins (UBJ) ARIMA modeling.

Carter, M. M. and Elsner, D.J.B., (1997), used results from a factor analysis regionalization of non-tropical storm convective rainfall over the island of Puerto Rico, a statistical methodology is investigated for its potential to forecast rain events over limited areas. Island regionalization is performed on a 15-yr dataset, while the predictive model is derived from 3 yr of surface and rainfall data. The work is an initial attempt at improving objective guidance for operational rainfall forecasting in Puerto Rico. Surface data from two first-order stations are used as input to a partially adaptive classification tree to predict the occurrence of heavy rain. Results from a case study show that the methodology has skill above climatology—the leading contender in such cases. The algorithm also achieves skill over persistence. Comparisons of forecast skill with a linear discriminant analysis suggest that classification trees are an easier and more natural way to handle this kind of forecast problem. Synthesis of results confirms the notion that despite the very local nature of tropical convection, synoptic-scale disturbances are responsible for prepping the environment for rainfall. Generalizations of the findings and a discussion of a more realistic forecast setting in which to apply the technology for improving tropical rainfall forecasts are given.

Makridakis et al. (1998) has given a decent account on exponential smoothing ways.

Al-Ansari et al. (2003) proscribed the applied math analysis of the rainfall measurements for 3 meteorological stations in Jordan: Amman aerodrome (central Jordan), Irbid (northern Jordan) and Mafraq (eastern Jordan). Traditional applied math and power spectrum analyses as well as ARIMA model were performed on the semi-permanent annual rainfall measurements at the 3 stations. The result shows that potential periodicities of the order of 2.3 - 3.45, 2.5 - 3.4 and 2.44-4.1 years for Amman, Irbid and Mafraq stations, respectively, were obtained. A statistic model for every station was adjusted, processed, diagnostically checked and finally an ARIMA model for every station is established with a ninety fifth confidence interval and also the model was used to forecast five years annual rainfall values for Amman, Irbid and Mafraq meteorological stations.

Al-Ansari, A., Al- Shamali B.and Shatnawi A.,(2006) used statistical Analysis of rain records at 3 major meteorological stations in Jordan, Al-Mararah University.

Ingsrisawang, L. et.al, (2010), made use of three statistical methods: First-order Markov Chain, Logistic model, and Generalized Estimating Equation (GEE) in modeling the rainfall prediction over the eastern part of Thailand. Two daily datasets during 2004-2008, so-called Meteor and GPCM, were obtained from Thai Meteorological Department (TMD) and Bureau of the Royal Rain Making and Agricultural Aviation (BRRAA). The Meteor observation consists of the average of rain volumes (AVG) from 15 local weather stations, and the observation of the Great Plain Cumulus Model (GPCM) includes 52 variables, for example, temperature, humidity, pressure, wind, atmospheric stability, seeding potential, rain making operation, and rain occurrence. Merging and matching between the GPCM dataset and Meteor observations, the GPCM+Meteor dataset was generated including 667 records with 66 variables. The first-order Markov chain model was then built using the Meteor dataset to predict two transitional probabilities of a day being wet given the previous day being wet or being dry, P(W/W) and P(W/D), respectively. The odds ratio(OR) was computed from these probabilities and gave the value of 6.85, which indicated that it was about 7 times more likely to be a wet day given the previous day was also wet within the eastern region of Thailand, than that given the previous day was dry. Next, the logistic models were also fitted using the Meteor dataset by taking account of cyclical effect in modeling for the prediction of P(W/W) and P(W/D), respectively. The models showed that the odds ratios of being wet days are not constant over day t during the years 2004-2008. Finally, the GEE method was applied with the GPCM+Meteor dataset to study the effects of weather conditions on the prediction of rainfall estimates on wet days, by taking account of correlation structure among observations. The variables of -15 ºc isotherm height and K-Index were shown statistically significant for the prediction of rainfall estimates at a 0.05 level. In order to effectively detect the rain conditions and make the right decisions in cloud-seeding operations, the statistical methods presented in this study can help in deriving the useful features from the rain and weather observations and modeling the rain occurrence.

Seyed et al.,(2011) used time series methodology to model weather parameter in Islamic Republic of Iran at Abadeh Station and counseled ARIMA(0,0,1)(1,1,1)12 because the best appropriate monthly rainfall information and ARIMA(2,1,0)(2,1,0)12 for monthly average temperature for Abadeh station.

Seyed, A., Shamsnia, M.,Naeem,S. and Ali, L.,(2011) modelled weather parameter using random methods(ARIMA Model)(Case Study:Abadeh station,Iran).

Mahsin et al. (2012) used Box-Jenkins methodology to create seasonal ARIMA model for monthly rainfall information taken for Dhaka station, Bangladesh, for the amount from 1981-2010. In their paper, ARIMA (0, 0, 1) (0, 1, 1)12 model was found adequate and also the model is employed for forecasting the monthly rain.

III. Data Mining Forecasting Techniques

There are several application domains of DM with time series data, being that one important application domain is time series data analysis for forecasting. Intelligent forecasting models have achieved better results than traditional methods, particularly in time series data analysis for forecasting. Methods based on computational intelligence techniques include such as neural networks (NN) or genetic algorithms (GA). Hybrid methods combining more than one technique are also commonly found in the literature. Computational intelligence methods for time series forecasting generally fall into two major categories: (i) Methods based on NN; and (ii) Methods based on evolutionary computation.

Neural networks are widely applied to model several of nonlinear hydrologic processes like weather forecasting. Additional elaborate discussion concerning the ideas and applications of ANN in geophysical science will be referred to within the 2 technical papers prepared by the ASCE Task Committee on Application of Artificial Neural Networks in geophysical science as appeared within the Journal of Hydrologic Engineering (ASCE, 2000).

Hu(1964) initiated the implementation of ANN, a very important soft computing methodology in weather forecasting. Since the previous few decades, ANN a voluminous development within the application field of ANN has unfolded new avenues to the forecasting task involving environment connected development.

French et al. (1992), took a pioneering work in applying ANN for rain forecasting, that used a neural network to forecast two-dimensional rainfall, 1 h prior to. Their ANN model used present rainfall information, generated by a mathematical rainfall simulation model, as an input data. This work is, however, restricted in a very range of aspects. For instance, there's a trade-off between the interaction and also the training time, that couldn't be simply balanced. The amount of hidden layers and hidden nodes appear short, compared with the amount of input and output nodes, to reserve the upper order relationship required for adequately abstracting the method. Still, it's been thought-about because the 1st contribution to ANN’s application and established a brand new trend in understanding and evaluating the roles of ANN in investigating complicated geophysical processes.

Michaelides et al (1995) compared the performance of ANN with multiple linear regressions in estimating missing rainfall information over Cyprus.

Kalogirou et al (1997) enforced ANN to reconstruct the rainfall over the time series over Cyprus.

Monica Adyal and Fred Collopy, (1998) identified eleven guidelines that could be used in evaluating this literature. Using these, they examined applications of NNs to business forecasting and prediction. They located 48 studies done between 1988 and 1994. For each, they evaluated how effectively the proposed technique was compared with alternatives (effectiveness of validation) and how well the technique was implemented (effectiveness of implementation). It was found that eleven of the studies were both effectively validated and implemented. Another eleven studies were effectively validated and produced positive results, even though there were some problems with respect to the quality of their NN implementations. Of these 22 studies, 18 supported the potential of NNs for forecasting and prediction.

Lee et al(1998) applied ANN in rainfall prediction by rending the offered information into same subpopulations. Wong et al (1999) made fuzzy rules bases with the help of Kyrgyzstani monetary unit and back-propagation neural networks and so with the assistance of the rule base developed predictive model for rainfall over Switzerland using spatial interpolation.

Koizumi (1999) used an ANN model using microwave radar, satellite and weather-station information along with numerical products generated by the Japan Meteorological Agency (JMA) and also the model was trained using 1-year information. It absolutely was found that the ANN skills were higher than the persistence forecast (after three h), the regression toward the mean forecasts, and also the numerical model precipitation prediction. Because the ANN model was trained with only 1 year information, the results were limited. The author believed that the performance of the neural network would be improved once additional training information became available. It's still unclear to what extent every predictor contributed to the forecast and to what extent recent observations may improve the forecast.

Toth et al. (2000) compared short-time rainfall prediction models for real-time flood forecasting. Completely different structures of auto-regressive moving average (ARMA) models, ANN and Nearest-Neighbors approaches were applied for forecasting storm rainfall occurring within the Sieve river basin, Italy, within the amount 1992- 1996 with lead times variable from 1 to 6 h. The ANN adaptative activity application proved to be stable for lead times longer than three hours, however inadequate for reproducing low rainfall events. Abraham et al. (2001) used an ANN with scaled conjugate gradient algorithmic rule (ANN-SCGA) and evolving fuzzy neural network (EfuNN) for predicting the rainfall time series. In the study, monthly rainfall was used as input data for training model. The authors analyzed eighty seven years of rainfall information in Kerala, a state within the southern a part of the Indian dry land. The empirical results showed that neuro-fuzzy systems were economical in terms of getting higher performance time and lower error rates five compared to the pure neural network approach. nevertheless, rainfall is one in all the twenty most complicated and tough parts of the geophysical science cycle to grasp and to model due to the tremendous range of variation over a wide range of scales both in space and time.

Pucheta Julian A, et. al , (2010), presented a feed-forward NN based NAR model for forecasting time series. The learning rule used to adjust the NN weights is based on the Levenberg-Marquardt method. The approach is tested over five time series obtained from samples of the Mackey-Glass delay differential equations and from monthly cumulative rainfall. Three sets of parameters for MG solution were used, whereas the monthly cumulative rainfall belongs to two different sites and times period, La Perla 1962-1971 and Santa Francisca 200-2010, both located at Córdoba, Argentina. The approach performance presented is shown by forecasting the 18 future values from each time series simulated by a Monte Carlo of 500 trials with fractional Gaussian noise to specify the variance. R. Adhikari and R.K.Agarwal, (2012), in their work comprehensively explores the outstanding ability of Artificial NEURAL network in recognizing and forecasting strong seasonal patterns without removing them from the raw data. Six real world time series data having dominant seasonal fluctuations are used in the present work. The emperical results show that the properly designed ANN;s are remarkably efficient in directly forecasting strong seasonal variation as well as outperform each of the three statistical models for all six time series.

IV. CONCLUSION

This paper presents a literature review of the use of data mining with time series data. This literature review is very useful, since it brings a better understanding of the field of study, and this is an important contribution of this paper.

From the literature review it can be concluded that this subject attracts a great deal of interest by researchers. However, several research issues remain unexplored. One of the ones that were identified during this research is related with the combined use of fundamental and technical issues.

Future research directions include the study of ways to select the best features for DM, with special reference to computational intelligence techniques with time series data analysis. The existence of features with different frequencies is a concern, and methods that will help how to envisage this problem will be made use of for future research work.

REFERENCES

[1] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). Time series analysis : Forecasting and control, Pearson Education, Delhi.

[2] Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting Methods and Applications, 3rd Edition, John Wiley, New York.

[3] Chiew, F.H.S., M.J. Stewardson and T.A. McMahon,1993. Comparison of six rainfall-runoff

[4] Kuo, J.T. and Y.H. Sun, 1993. An interventionmodel for average 10 day stream flow forecast andsynthesis. J. Hydrol., 151: 35-56.

[5] Langu, E.M., 1993. Detection of changes in rainfalland runoff patterns. J. Hydrol., 147: 153-167

[6] M.J.C., Hu, Application of ADALINE system to weather forecasting, Technical Report, Stanford Electron, 1964

[7] Michaelides, S. C., Neocleous, C. C. & Schizas, C. N. “Artificial neural networks and multiple linear regression in estimating missing rainfall data.” In: Proceedings of the DSP95 International Conference on Digital Signal Processing, Limassol, Cyprus. pp 668–673, 1995.

[8] Kalogirou, S. A., Neocleous, C., Constantinos, C. N., Michaelides, S. C.& Schizas, C. N.,”A time series construction of precipitation records using artificial neural networks. In: Proceedings of EUFIT ’97 Conference, 8–11 September, Aachen, Germany. pp 2409–2413 1997.

[9] Lee, S., Cho, S.& Wong, P.M.,”Rainfall prediction using artificial neural network.“,J. Geog. Inf. Decision Anal. 2, 233–242 1998.

[10] Wong, K. W., Wong, P. M., Gedeon, T. D. & Fung, C. C. , “Rainfall Prediction Using Neural Fuzzy Technique.” 1999

[11] E.Toth, A.Brath, A.Montanari,” Comparison of short-term rainfall prediction models for real-time flood forecasting”, Journal of Hydrology 239 (2000) 132–147

[12] Koizumi, K.: “An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network”, Weather Forecast., 14, 109–118, 1999.

[13] Ajith Abraham, Dan Steinberg and Ninan Sajeeth Philip,” Rainfall Forecasting Using Soft Computing Models and Multivariate Adaptive Regression Splines”, 2001.

[14] French, M. N., Krajewski, W. F., and Cuykendall, R. R.: Rainfall forecasting in space and time using neural network, J. Hydrol., 137, 1–31, 1992.

[15] Koizumi, K.: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network, Weather Forecast., 14, 109–118, 1999.

[16] Toth, E., Montanari, A., and Brath, A.: Comparison of short-term rainfall prediction model for real-time flood forecasting, J. Hydrol., 239, 132–147, 2000.

[17] Maier, R. H. and Dandy, G. C.: The use of artificial neural network for the prediction of water quality parameters, Water Resour. Res., 32(4), 1013–1022, 1996.

[18] Maier, R. H. and Dandy, G. C.: Comparison of various methods for training feed-forward neural network for salinity forecasting, Water Resour. Res., 35(8), 2591–2596, 1999.

[19] ASCE: Task Committee on Application of Artificial Neural Networks in Hydrology. I: Preliminary Concepts, J. Hydrol. Eng., 5(2), 115–123, 2000.

[20] ASCE: Task Committee on Application of Artificial Neural Networks in Hydrology. II: Hydrologic Applications, J. Hydrol. Eng., 5(2), 124–137, 2000

[21] Pucheta Julian A, et. al , “A Feed-Forward Neural Networks-Based Nonlinear Autoregressive Model for Forecasting Time Series”, Computación y Sistemas Vol. 14 No. 4, pp 423-435, 2010.

[22] Adhikari, R., et. al., “Forecasting strong seasonal time series with Artificial Neural Network”, Journal of Scientific and Industrial Research, Vol. 71, pp. 657-666, 2012.

[23] Adya1, M and Collopy, F., “How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation”, Journal of Forecasting, J. Forecast. 17, 481±495, 1998.

[24] Carter, M. M. and Elsner, D.J.B., “A Statistical Method for Forecasting Rainfall over Puerto Rico”, American Meteorological Society, vol. 12, pp.515-525, 1997.

[25] Ingsrisawang, L. et.al., “Applications of Statistical Methods for Rainfall Prediction over the Eastern Thailand”, Proceedings of the MultiConference of Engineersand Computer Scientists, vol. III, IMECS 2010, March 17-19, 2010, Hong Kong

[26] Bhaskaran, S., “Time Series Data Analysis for long term forecasting and scheduling of organizational resources – few cases”, International Journal of Computer Applications (0975 – 8887) Volume 41– No.12, March 2012

[27] Winstanley, D.,(1973a). “Recent rainfall trends in africa, the middle East and India”, Nature. 243: 464–465pp.

[28] Winstanley, D.,(1973b). “Rain Patterns and General atmospherical circulation’, Nature. 245: 190–194pp.

[29] Laban, A.J.and Ogallo,H., (1986), “Stochastic modelling of regional annual rain anomalies in East Africa”, Journal of Applied Statistics, Vol.13.

[30] Harvey, R., Andrew C. and Souza, R.C., (1987), “Assessing and Modeling the alternate Behavior of rainfall in North-East Brazil”, Journal of Climate and Applied Meteorology, Vol.26, 1339-1344pp.

[31] Mahsin, M.D. Yesmin, A. and Monira, B.,(2012). Modeling rain in Dacca (national capital} Division of Bangladesh using Time SeriesAnalysis. Journal of Mathematical Modelling and Application, Vol. 1, No.5, 67-73pp.