August 2016

Improving efficacy of library Services: ARIMA modelling for predicting book borrowing for optimizing resource utilization

K Kumar
kumarkkutty@gmail.com

M M.Alpha Raj

Improving efficacy of library Services: ARIMA modelling for predicting book borrowing for optimizing resource utilization

K. Kumar[a] and M.Alpha Raj[b]

[a]Assistant Professor, Library and Information Science, College of Veterinary Science, Sri Venkateswara Veterinary University, Proddatur – 516 360, Email: kumarkkutty@gmail.com

[b]Assistant Professor, Dept. of Pharmacology and Toxicology, College of Veterinary Science, Sri Venkateswara Veterinary University, Proddatur – 516 360, Email: alpharajm@gmail.com

Abstract

Book borrowing is a key service in libraries. Library users frequently visit the library for borrowing books compared to other library services. To predict book-borrowing service in a college library, Auto Regressive Integrated Moving Average (ARIMA) model has been developed from the data pertaining to book borrowing during the year 1998 to 2013. The study found that the number of books borrowed one month and twelve months earlier could estimate the number of books borrowed in a month. The study used a fitted model for predicting book borrowing for the year 2014 by two alternative approaches: 12-steps ahead versus 1-step ahead. The calculations show that there was no significant difference (P=0.928; Wilcoxon signed rank test) between 1-step and 12-steps ahead approach for predicting book borrowing. However, the Root Mean Squared Error (RMSE) in 1-step ahead approach (109.57) was lower than 12-steps approach (131.33). The study findings indicate that ARIMA models are useful for monitoring book borrowing in institutional libraries. Furthermore, these models can predict library usage trends.

Keywords: ARIMA Modelling; Book borrowing; Library; Prediction

Introduction

Academic libraries are an essential component of colleges and universities throughout the world and play a central role in academic life[1]. Regular library usage by students improves critical thinking[2, 3], student learning and engagement, collaborative learning, student-faculty interactions, and academic challenges[4, 5]. Even the academic performance of students improves with library usage[6,7]. Academic librarians face immense pressure to meet the needs of ever-increasing users and commitments for modernization and improvements[8]. Further, with the advent of information-communication revolution, academic libraries are exploring service developments to support a series of new scenarios such as publication and scholarly communications; extensive use of digital resources; increasing heterogeneous student population; ICT-based learning and distance learning. Hence, libraries with limited resources, particularly in terms of staffing face enormous challenges in management[9]. Librarians need to predict library usage over time, as this helps in planning the resources required to meet the usage. The future usage might not be stationary and there could be periods of huge demand, interleaved with long periods. A model for predicting the library usage should take into account this variability to be accurate in its prediction. Hence, seasonal prediction models are likely to improve the accuracy of prediction. The study proposes a prediction model that forecasts library usage for a period of 12 months.

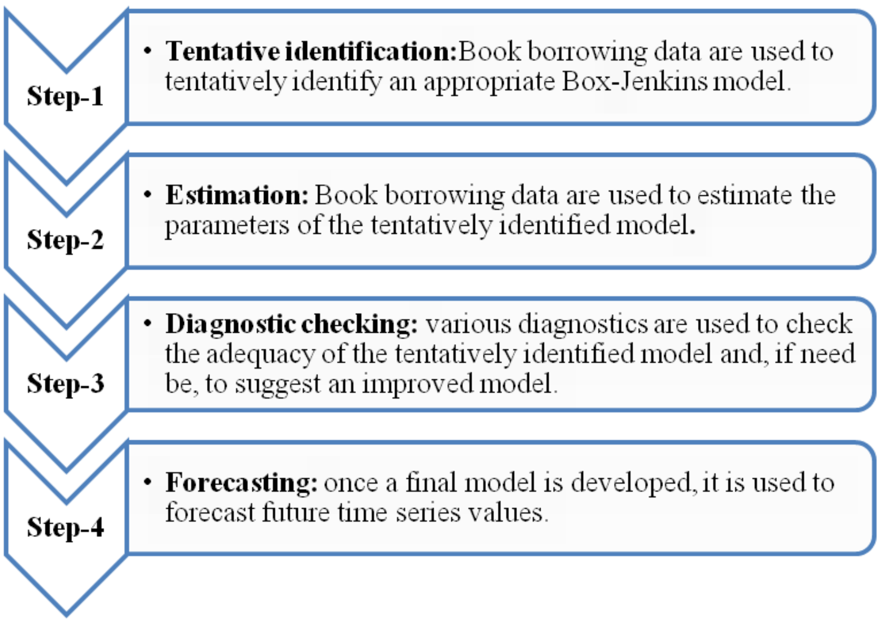
Time series analysis allows examining temporal dependency of data and generating predictions. Time series consists of two components: trend and seasonality. Trend depicts a linear or non-linear component that is non-repetitive within the time range, whereas seasonality repeats at systematic intervals over time[10]. Time series analysis using Autoregressive Integrated Moving Average (ARIMA) models are very effective in modelling the time-dependent data as they structure time dependence between the observations[11]. The predictions made using ARIMA models are more accurate than those obtained by other statistical methods[12-15]. Earlier, ARIMA models are successfully forecasted household electric consumption[16]; software evolution[17]; incidence of several infectious diseases[12,18-25] and use of health facilities[26,27]. However, such predictions are not available in the context of library usage. Hence, in this study, the authors used Box-Jenkings ARIMA modelling to monitor and predict book borrowing in Sri Venkateswara College of Engineering and Technology, Chittoor. The data pertaining to book borrowing from 1998 to 2013 fit into an ARIMA model using Box-Jenkings approach[28]. The fitted model has been used to predict book borrowing for the year 2014 with 12-steps ahead and 1-step ahead approach. Root mean squared error (RMSE) was used to compare the predictive power of the two approaches. Further, both approaches have been tested using Wilcoxon signed rank test[29].

Objectives of the study

- To detect any seasonal pattern in the book borrowing services and identify the underlying factors associated with it;
- To identify trends in the distribution of book borrowing over time; and
- To validate the usage of ARIMA model with seasonal information to accurate forecasting of book borrowing.

Methodology

The study was conducted at the college library of Sri Venkateswara College of Engineering and Technology, Chittoor established in 1998. For the analyses, monthly data on book borrowing for the period 1998–2013 was used to fit an ARIMA model, which later used to predict out-of-sample book borrowing for the year 2014. Since estimation procedures are available only for stationary series, Augmented Dickey-Fuller (ADF) unit root test and Box-Jenkins approach on ARIMA modelling of time series, used to check the stationary series of the data, which was ascertained by observing the plot of book borrowing that involves a four-step process[11, 15,24,28-33].



Based on the correlogram of Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF), the order of non-seasonal (p, d, q) and seasonal (P, D, Q) component; Auto Regressive (AR) parameters (p and P); Moving Average (MA) parameters (q and Q); the need for non-seasonal and seasonal differencing (d and D) [34,35] determined. The goodness of fit of the model validated using Bayesian Information Criterion (BIC) and Ljung-Box statistics. In the identified ARIMA model, two approaches employed to estimate the values for the year 2014, in which out of sample predicted values with 95% prediction intervals are calculated.

- The first approach was to predict the subsequent 12-steps ahead for the entire year 2014 and second to predict 1-step ahead, that is, January 2014 only.

- Then, as the observed value for January 2014 obtained, the authors included the same in the model and re-estimated the ARIMA model, and predicted next1-step, that is February 2014, this process continued until the end of the year 2014.

The predictive power of both approaches by RMSE accessed, which measures the difference between fitted and observed values. Lower RMSE indicates increased predictive power. Further, the analysis applied Wilcoxon signed-rank test to assess the statistical difference in the errors between both approaches[15, 29]. ADF unit root test and Wilcoxon signed rank test carried out in Microsoft Excel 2013 and ARIMA modelling in the Statistical Package for Social Sciences (IBM SPSS 19.0 V). Expert modeller of SPSS used for 1-step ahead predictions.

Analysis

The box plot of book borrowing showed no trend and the data remained stationary (Figure 1). In addition, ADF test (statistic = -4.75; p=0.01; lag order = 5) showed that the data was stationary without significant trend. Correlogram of sample plots of ACF and PACF indicated that non-seasonal and seasonal parameters are required in the model (Figure 2 and 3). The sinusoidal exponential decay in ACF associated with an oscillating PACF suggested that the model should contain a non-seasonal AR and MA components. In the time lag (lag=l2) correlogram, the ACF showed exponential decay, whereas PACF showed spikes at lags one and two, and later no significant spikes suggesting a seasonal AR (P =2) (Figure 4 and 5). After analysis, the best model was found to be a multiplicative ARIMA $(1,0,1) \times (2,0,0)_{12}$ (Normalized BIC=8.898,Table1) on the non-differenced book borrowing data (multiplicative ARMA$[2,1]\times[2,0]_{12}$, as no differencing required). After developing the model, the plots of residuals of ACF and PACF showed no remaining temporal correlation (Figure 6) indicating that, the model is a good fit. Higher seasonal ARIMA models like (2,0,1) (2,0,0), (2,0,2) (2,0,0) and (1,0,2) (2,0,0) also fitted the data, but they were rejected based on Occam's razor-among competing models that predict equally well, the one with the least complexity should be selected.

Seasonal ARIMA (1,0,1) $(2,0,0)_{12}$ model equation (Table 2)

$$\hat{Y}_t = Y_t - \varphi_1 Y_{t-1} + \Phi_1 Y_{t-12} + \Phi_1 \varphi_1 Y_{t-13} + \Phi_2 Y_{t-24} + \Phi_2 \varphi_1 Y_{t-25} + \theta_1 Y_{t-1} - \varphi_1 \theta_1 Y_{t-2} - \Phi_1 \theta_1 Y_{t-13} + \Phi_1 \varphi_1 \theta_1 Y_{t-14}$$

Where, $\hat{Y}_t$ =Predicted value; Yt = Constant

φ_1 =Non-seasonal Auto Regression coefficient

θ_1 = Non-seasonal Moving Average coefficient

Φ_1, Φ_2= Seasonal Auto Régression coefficients

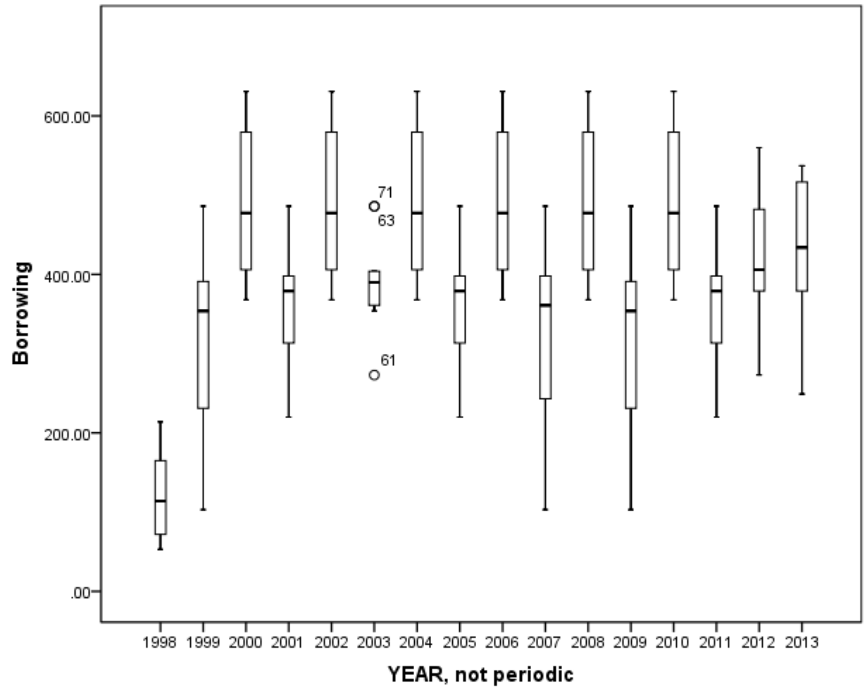


Fig. 1-- Box plot of book borrowing cases from the year 1998 to 2013 presenting no particular trend

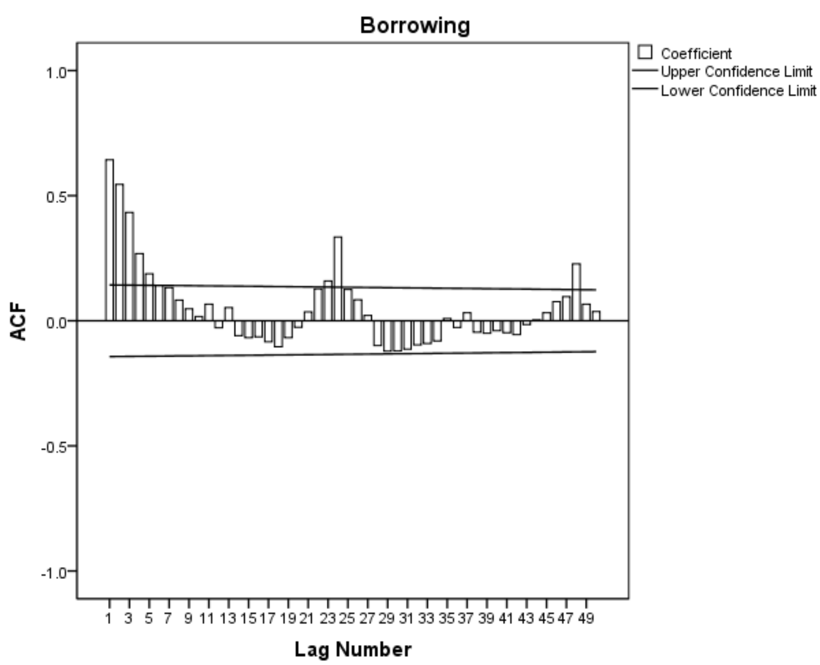


Fig. 2-- ACF shows a sigmoid curve with exponential decay

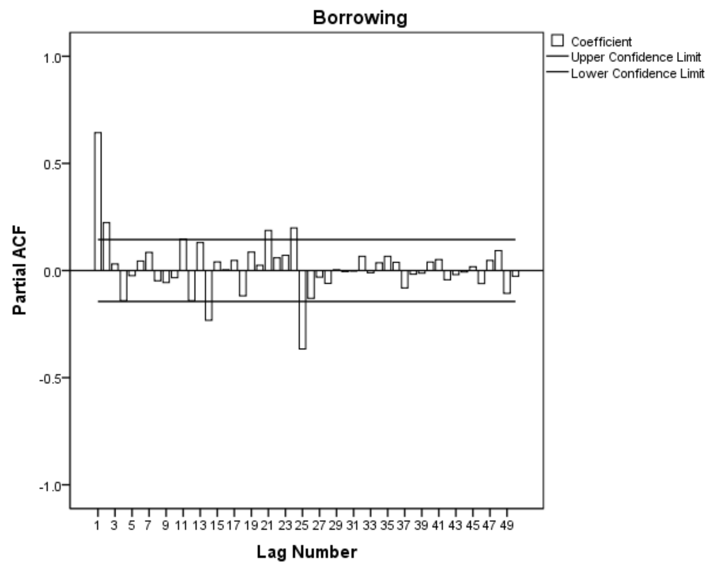


Fig. 3-- PACF showing multiple peaks until lag25 and then cutting off

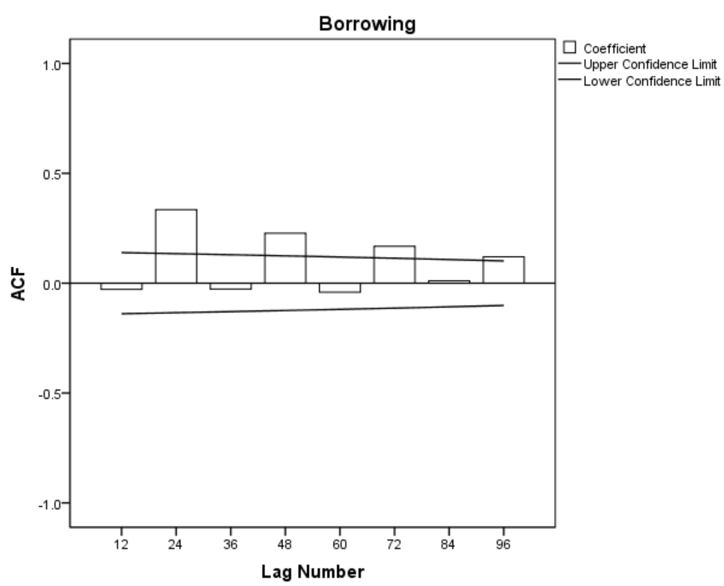


Fig. 4-- ACF at periodic lag of 12 indicating exponential decaying

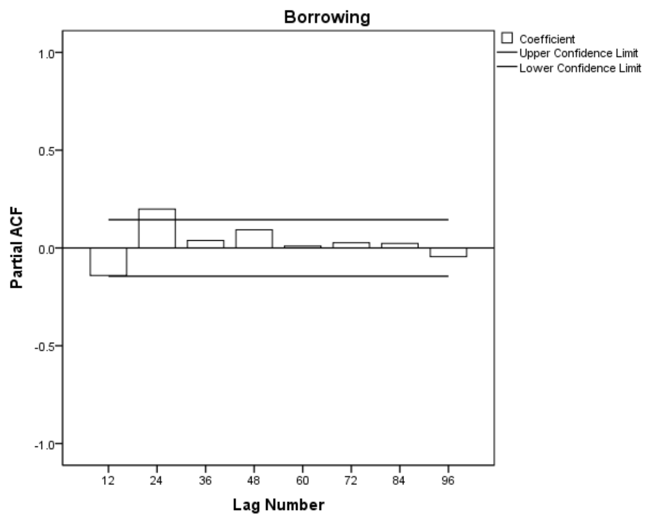


Fig. 5-- PACF at a periodic lag of 12 indicating spikes at lags 1 and 2 and cutting off later

In addition, Ljung–Box test statistic was calculated, which supported the independence of the time series residuals (Statistic=12.81;df=15; P=0.617). The model fitted (1998–2013) and predicted values (the year 2014) followed the observed book borrowing in the same year (Figure 7). Graphically, the predicted values estimated using the 12-step ahead approach showed less dispersion from observed values (RMSE=139.87) than 12-step ahead approach (RMSE=140.30) (Figure 8). The Wilcoxon signed rank test statistic "W" is simply the smallest of the rank totals. The smaller it is then the less likely it is to have occurred by chance. Wilcoxon signed rank test showed that the predictions of 12-step ahead approach were significantly different than1-step ahead approach (W= 9; Critical value = 10).

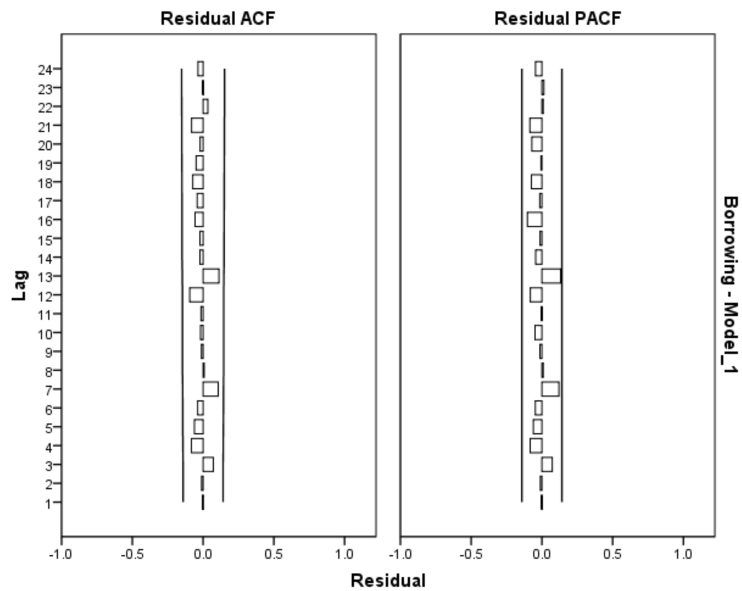


Fig. 6--Residual ACF and PACF plots after modelling illustrate no remaining temporal correlation

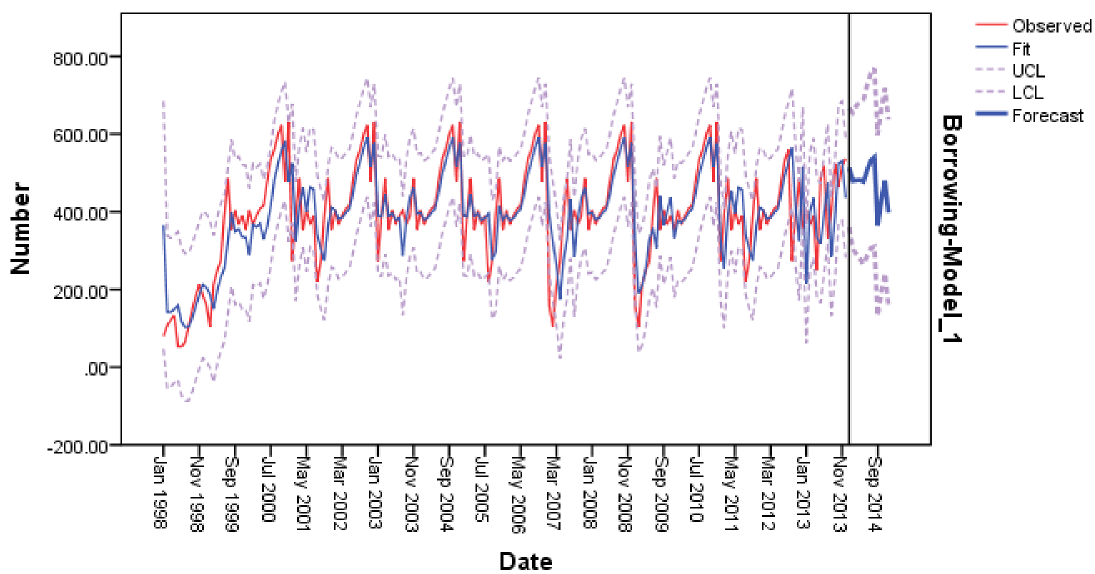


Fig. 7-- Plot of the observed and forecasted book borrowing

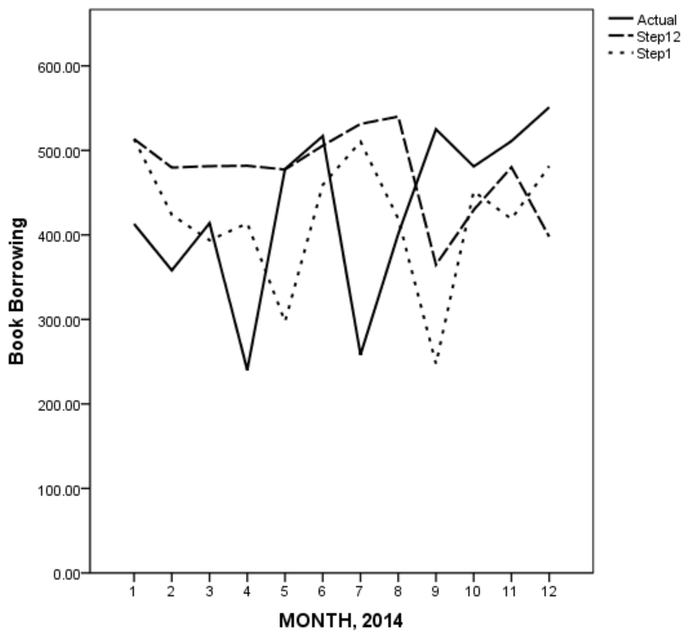


Fig. 8--Plot of actual book borrowing Vs 12-step ahead and 1-step ahead approach predictions

Table 1--ARIMA (1, 0, 1) (2, 0, 0)$_{12}$ model statistics		
Sl. No.	Statistic	Value
1	Stationary R-squared	0.629
2	R-squared	0.629
3	Root Mean Squared Error (RMSE)	80.978
4	Mean Absolute Percentage Error (MAPE)	20.725
5	Mean Absolute Error (MAE)	57.568
6	Maximum Absolute Percentage Error	357.159
7	Maximum Absolute Error	291.353
8	Normalized Bayesian Information Criterion	8.898

Table 2- ARIMA model (1, 0, 1) (2, 0, 0)$_{12}$ parameters					
Parameter	Lag No.	Estimate	Standard Error	T-value	Significance
Constant		356.727	74.41	4.92	0.000
AR	Lag 1	0.900	0.037	24.64	0.000
MA	Lag 1	0.346	0.084	4.120	0.000
AR, Seasonal	Lag 2	0.590	0.061	9.724	0.000

Discussion and Conclusion

Book borrowing is an important aspect of libraries[36] and attracts many people to the library. Academic libraries provide a unique opportunity for students to borrow books free of cost. This often benefits students who generally have limited financial resources and cannot afford to buy high priced books. Hence, they prefer to borrow books rather than to buy them,[37] at the same time, book lending is a challenging aspect of librarians requiring more resources to meet the demand. Therefore, prior information on the number of book borrowings in a month will help the librarian to plan and prepare for the task effectively. Various techniques proposed[38] to forecast a given time series; studies show that models from the ARIMA successfully used, as well as regression approaches based on e.g. linear, non-linear regression, neural networks, and Support Vector Regression. The difference lies in forecasting methodology. The application to plenty of time series underlines its usefulness.

In this study, an ARIMA model developed, that closely fits book borrowing in the college library of SVCET, Chittoor. The Auto Regressive and Moving Average parameters of the model entail the number of book borrowing cases in a month can be estimated by previous cases occurring one or twelve months prior; found from the analysis that 12-step ahead approach better agrees with the observed book borrowing and was significantly different ($P<0.05$) from 1-step ahead prediction. The model indicated that peak book borrowing occurs in August (539.99) [310.43-769.55] and least in September (364.72) [131.62-597.81). However, an increasing trend in book borrowing observed starting from June (505.69) [286.21-725.17] through August (539.99) [310.43-769.55]. This increase in trend is a result of semester examinations, scheduled during the months of June and July every year. Similarly, the actual book borrowing data showed peak borrowing during the month of June (517).

ARIMA models of time series applied to a variety of library functions including bibliometric series,[39] electronic databases[40] if the numerical data can be obtained. This procedure also used to evaluate predictive scenarios for search engine transactional[41] logs, resource allocation, and usage for server[42] monthly circulation[43,44] of library books, and more. The study concludes that ARIMA modelling successfully predicted book borrowing in 12-step ahead approaches. However, forecasting requires the presence of ‘cause and effect’ relationship successfully predict into the future. In this study, book borrowing in academic library appears directly related to the examination schedule. Hence, these predictions used successfully deploy working staff in libraries. Moreover, such predictions are also helpful in making a strong case for increasing budgetary allocations and workforce for libraries.

References

1. Kuh GD and Gonyea RM, The Role of the Academic Library in Promoting Student Engagement in Learning, *College & Research Libraries*, 64 (4) (2003) 256-282.
2. Ory JC and Braskamp LA, Involvement and Growth of Students in Three Academic Programs, *Research in Higher Education* 28 (2) (1988) 116–129.
3. Whitmire E, Development of Critical Thinking Skills: An Analysis of Academic Library Experiences and Other Measures, *College & Research Libraries*, 59 (3) (1998) 266–273.
4. Laird T F N and Kuh G D, Student Experiences with Information Technology and Their Relationship to Other Aspects of Student Engagement, *Research in Higher Education*, 46 (2) (2005) 211–233.
5. Pike, G R, Smart J C, Kuh G D and Hayek J C, Educational Expenditures and Student Engagement: When Does Money Matter?, *Research in Higher Education*, 47 (7) (2006) 847–872.
6. Barkey P, Patterns of Student Use of a College Library, *College and Research Libraries*, 26 (2) (1965) 115–118.
7. Knapp P B, *The Monteith College Library Experiment*, (Scarecrow Press; New Delhi), 1966.
8. Keeling RP, Wall A F, Underhile R and Dungy G J, *Assessment Reconsidered: Institutional Effectiveness for Student Success*, 1stedn, (NASPA; Washington), 2008.
9. Brophy P, *The Library in the Twenty-first Century: new services for the information age*, (Library Association Publishing; London), 2001.
10. Hill T and Lewicki P, *Statistics: Methods and Applications*, (StatSoft Inc; Oklahoma), 2007.
11. Helfenstein U, The use of transfer function models, intervention analysis and related time

series methods in epidemiology. *International Journal of Epidemiology,* 20 (3) (1991) 808–815.

12. Choi K and Thacker SB, An evaluation of influenza mortality surveillance, 1962–1979. 1. Time-series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology,* 113 (3) (1981) 215-226.
13. Farmer RD and Emami J, Models for forecasting hospital bed requirements in the acute sector, *Journal of Epidemiology and Community Health*, 44 (4) (1990) 307–312.
14. Milner PC, Ten-year follow-up of ARIMA forecasts of attendances at accident and emergency departments in the Trent region, *Statistics in Medicine*, 16 (18) (1997) 2117–2125.
15. Nobre FF, Monteiro AB, Telles PR and Williamson G D, Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine*, 20 (20) (2001) 3051–3069.
16. Chujai P, Kerdprasop N and Kerdprasop K, *Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models*, In: Proceedings of the International Multi-Conference of Engineers and Computer Scientists, IMECS, Hong Kong 13-15 March, 2013.
17. Goul M, Fonte N, Wermelinger M and Brito e Abreu, F. *Software evolution prediction using seasonal time analysis: a comparative study*. In: Proceedings of 16th European Conference on Software Maintenance and Reengineering, Szeged, Hungary 27-30 March 2012.
18. Musa MI, Malaria disease distribution in Sudan using time series ARIMA model, *International Journal of Public Health Science*, 4(1) (2015) 7-16.
19. Liu Q, Liu X, Jiang B and Yang W, Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model, *BMC Infectious Disease*, 11 (2011) 218-224.
20. Jin R,Qiu H,Zhou X, Huang P, Wang Z and Wei J, Forecasting incidence of intestinal infectious diseases in mainland China with ARIMA model and GM(1,1) model, *Fudan University Journal of Medical Sciences*, 35(5) (2008) 675-0.
21. Trottier H, Philippe P and Roy R, Stochastic modeling of empirical time series of childhood infectious diseases data before and after mass vaccination, *Emerging Themes in Epidemiology,* 3 (2006) 9.

22. Tong S and Hu W, Climate variation and incidence of Ross River virus in Cairns, Australia: a time-series analysis, *Environmental Health Perspectives*, 109 (12) (2001) 1271–1273.

23. Allard R, Use of time-series analysis in infectious disease surveillance, *Bulletin of World Health Organization*, 76 (4) (1998) 327–333.

24. Helfenstein U, Box-Jenkins modelling of some viral infectious diseases, *Statistics in Medicine*, 5 (1) (1986) 37–47.

25. Bowie C and Prothero D, Finding causes of seasonal diseases using time series analysis, *International Journal of Epidemiology*, 10 (1) (1981) 87–92.

26. Reis BY and Mandl KD, Time series modelling for syndromic surveillance. *BMC Medical Informatics & Decision Making*, 3 (2003) 2.

27. Earnest A, Chen MI, Ng D and Sin LY, Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research*, 5 (2005) 36.

28. Box GEP, Jenkins GM, Reinsel G C, *Time Series Analysis: Forecasting and Control*, 4th edn, (Wiley; San Francisco), 2008.

29. Heiberger RM and Holland B, *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-plus, R, and SAS*, (Springer; New York), 2004.

30. Heiberger RM and Teles P, Displays for direct comparison of ARIMA models. *The American Statistician*, 56 (2) (2002) 131–138.

31. Ljung G M and Box GEP, Measure of lack of fit in time-series models, *Biometrika*, 65 (2) (1978) 297–303.

32. Cheung YW and Lai KS, Lag order and critical values of the augmented dickey-fuller test. *Journal of Business and Economic Statistics*, 13 (3) (1995) 277-280.

33. Anderson S R, Auquier A,Hauck W W,Oakes D, Vandaele Wand Weisberg H I, *Statistical methods for comparative studies: techniques for bias reduction*, 1st edn, (John Wiley& Sons; USA), 1980.

34. McCleary R and Hay R, *Applied time series analysis for the social sciences*, (SAGE Publications; Beverly Hills), 1980.

35. ZickuhrK, Rainie L and Purcell K, *Library Services in the Digital Age. Pew Research Center's Internet & American Life Project*, (2013). Available at http://libraries.pewinternet.org/ files/legacy-pdf/PIP_Library%20services_Report.pdf

(Accessed on 05 May 2015).

36. Library & Information Commission, *Reading the Situation: Book reading, buying & borrowing habits in Britain, Library & Information Commission Research Report*, (Book Marketing Limited; London), 2000.

37. Udoumoh CN and Okoro CC, The Effect of Library Policies on Overdue Materials in University Libraries in the South-South Zone, Nigeria, *Library Philosophy and Practice*, (2007) 1-7.

38. Guajardo, José, Weber, Richard and Miranda, Jaime (2006) A Forecasting Methodology Using Support Vector Regression and Dynamic Feature Selection. *Journal of Information & Knowledge Management*. Vol. 5 (4). p329-335.

39. Murphy, L. J., Nelson, M. J., & Tague, J. M. (1975). Arima forecasts in information studies. In Husbands, Charles W., Ed.; Tighe, Ruth L., Ed. *Information Revolution. Proceedings Of The 38Th American Society For Information Science Annual Meeting*, Boston, Massachusetts, Volume 12.

40.Asha, B. (2008). Stochastic model for the Electronic Access of an Article. *SRELS Journal of Information Management*, 45(4).pp 391-398.

41. Zhang, Ying; Jansen, Bernard J.; Spink, Amanda (2009). Time series analysis of a Web search engine transaction log *Information Processing & Management*. Vol. 45(2). Pp.230-245.

42. Prangchumpol, Dulyawit; Sophatsathit, Peraphon; Lursinsap, Chidchanok; Tantasanawong, Panjai (2015).Resource Allocation with Exponential Model Prediction for Server Virtualization. *Journal of Digital Information Management*. Vol. 13 (5).pp385-398.

43. Brooks, T A (1984).Naive vs. sophisticated methods of forecasting public library circulations *Periodical Library & Information Science Research*. Vol. 6(2). Pp.205-214.

44. David A. Dickey, The book Bot Ultimatum! (2015) NC State University, John Vickery NC State University. Paper 1579-2015 Available at:
http://support.sas.com/resources/papers/proceedings15/1579-2015.pdf retrieved 20.02.2016