© 2015 IJIRT | Volume 1 Issue 12 | ISSN: 2349-6002

Efficient Forecasting of HTTP workload using Seasonal ARIMA Model in IaaS

Shital Makwana
Computer Engineering Department,
Silver oak college of Engineering & Technology,
Gujarat Technological University, India

***Abstract* — This paper aims at improving Quality of Service (QoS) in Infrastructure as a Service (IaaS) model in cloud computing by forecasting resource provisioning on demand basis. Quality of Service in IaaS model can be improved by provisioning of resources in such a way that it not only complies with Service Level Agreement (SLA) but also allocates resources in right amount so that efficient utilization of resources occurs leading to low cost. In a dynamic workload environment like web server, the QoS parameters viz. response time, throughput etc can be achieved only by elastic provisioning of resources. In this paper, seasonal auto-regressive integrated moving average (SARIMA) model is implemented to predict the future workload. Based on the forecast from this model, resources may be allocated to the cloud users which guarantee the minimum SLA violations.**

***Index Terms*—Cloud Computing, Quality of Service (QoS), Infrastructure as a Service (IaaS), Elasticity.**

I. INTRODUCTION

A *cloud* is a distinct IT environment which is designed for the aim of remotely provisioning scalable and measured IT resources. NIST defines cloud computing as "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [2]. Cloud provides ubiquitous access as services are available online so you get service wherever internet is available .different physical and virtual resources are dynamically assigned to multiple consumers on demand. As per application need resources are provisioned and released manually or automatically. This elasticity feature distinguishes cloud computing with other technology. Users are charged in pay-per-use model.

II. CLOUD SERVICE MODELS

There are basically three Service Models of Cloud computing.

Software as a Service (SaaS)

SaaS is software that is owned, delivered and managed remotely. Cloud provider maintains and manages the software services which are used by the cloud consumer.

Platform as a Service (PaaS)

PaaS provides ready to use environment for users to support entire lifecycle of custom application. Developers can write their applications according to the specifications of a particular platform.

Infrastructure as a Service (IaaS)

IaaS offers computing resources such as storage or processing which can be obtained as a service without requiring any physical hardware on their own site. It is the base layer for cloud computing which basically deals with virtual machines, storage, servers, networks, load balancers, and the IaaS cloud providers supply these resources on-demand [3]. Hardware cost for organization can be greatly reduced here. Amazon Web Service (AWS), Rackspace, Windows Azure etc provide Infrastructure as a Service.

III. CLOUD DEPLOYMENT MODELS

Cloud is generally classified by the owner of data center, whether it is owned by some third party, organization, community or combination of any of this. There are four deployment models.

Public cloud

A Public cloud is owned and managed by third parties. It is generally on premises of cloud provider. Public Cloud is provisioned for open use by the general public on pay per use basis.

Private cloud

A Private Cloud is fully owned by a single company who has total control over the applications run on the infrastructure, the place where they run, and the people or organizations using it [4].

Community cloud

Different communities with same objectives come together to form a cloud so it is available to specific group of people.

Hybrid cloud

It is made by composition of two or more infrastructures like public, private or community cloud. User can use private or community cloud and can expand to public cloud as needed.

INTRODUCTION TO TIME SERIES COMPONENTS

Time series is a collection of data points measured over a fixed interval of time. The data points may refer to any observable quantity such as daily closing price of a stock of a company, monthly population of a country, etc. In this work, the data points in the time series denote the hourly workload of a given application in cloud. Time series analysis can be useful, for example, to forecast the future data points, classification, clustering, etc. depending on the type of application. In the context of this thesis, time series analysis is used for forecasting of the workload and that of corresponding resources required to sustain the same, in order to take pro-active scheduling decisions. Time-series aims to identify the pattern in the data to make meaningful predictions. Formally, time series is defined in terms of random variables.[10]

ARIMA AND SEASONAL ARIMA MODEL FOR NON-STATIONARY SERIES

Most of the real World Series is not stationary. This is because of the presence of trend factor, seasonality, change in variance with time, etc. in the series. ARMA process assumes that the time series being analyzed is stationary. However, in case of non-stationary time series, the model cannot be directly applied. Trend and seasonality changes the mean with time, hence making the series non-stationary. This can be addressed by doing some transformations and operations on the series and making it stationary. For example, if the series has a linear trend, then differencing a series will remove the trend from series.

There is another important characteristic that the real-world time series exhibit, which is seasonality. Seasonality is apparent in number of human behavior related data. For example, organizational data based on day and night patterns which is typically dictated by the working hours of an organization usually has seasonal component. Similarly, based on business cycles of an organization, monthly or quarterly patterns are quite observable. However, the presence of seasonality in the time series also makes it non-stationary. In case of seasonal time series, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag S. Seasonal ARIMA is an extension of ARIMA model to include seasonal components, which capture the seasonal variations in the series.

SEASONAL ARIMA FOR WORKLOAD FORECASTING

Seasonal Auto-Regressive Integrated Moving Average (ARIMA) model is a five step iterative procedure which is able to forecast a broad class of random processes. The steps include stationarity checking and differencing, model identification, parameter estimation, forecasting and diagnostic check.

The seasonal autoregressive integrated moving average ($SARIMA$) model, $SARIMA(p,d,q)$ x $(P,D,Q)_s$ is given by

$$\Phi_p(B^s)\phi_p(B)(1-B^s)^D(1-B)^d X_t = \Theta_Q(B^s)\theta_q(B)\omega_t \quad \text{(Equation 1.)}$$

where, $\Phi_p(B^s)$ and $\Theta_Q(B^s)$ are the AR and MA polynomials of seasonal part, and $(1-B^s)$ is the difference operator with a seasonal difference lag S.

The http workload is represented in the form of time series and various operations are performed on the time series to forecast the future data. In present work, hourly workload is computed and presented in the form of time series. The workload for 20 days (480 hours) was used as training data for parameter estimation of seasonal ARIMA. The workload for next 7 days was forecasted and compared with the actual workload to find out the accuracy of the model. Step by step details of the method are as follows:

Stationarity checking and differencing

A time series is said to be wide sense stationary if its statistical mean and variance don't change with time. Also the auto-covariance of the series is dependent only on the time difference irrespective of the time of observation chosen. ARMA models can only forecast stationary time series so to forecast non-stationary time series one needs to convert it to stationary time series. A time series may have a seasonal component, trend or random components. Stationarity can be achieved by differencing the series if the series contains seasonal component. To identify the seasonal component, one can plot the data or compute the autocorrelation (ACF) and partial autocorrelation functions (PACF). The http workload of Kennedy Space center for twety days is plotted in fig.1.

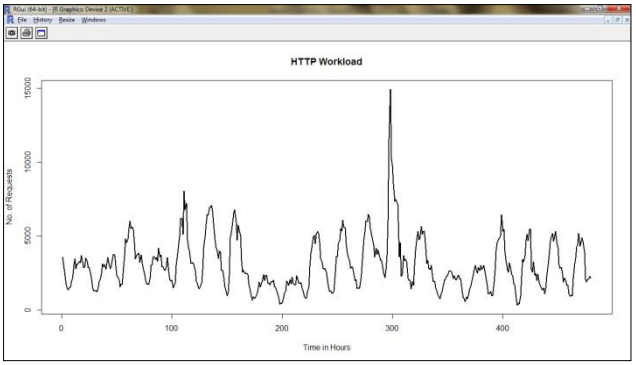

Fig. 1. Http workload

The first observation is that this request rate series is not stationary as the mean doesn't remain constant. This is because it has a seasonal trend of 24 hours. Clearly the workload has crest and troughs every 24 hours hence it has a seasonal component of 24 hours. The Augmented Dicky Fuller (ADF) test [15] shows whether the series is stationary or not. A value less than -3.5 shows that there is 95% confidence that the series is stationary. Although the ADF value (-7.6) shows the series to be stationary (-3.5 for 95% confidence) the ACF curve (Fig. 2) shows that a model

with large number of parameters will be required to forecast the time series since the order of moving average (MA) is equal to the lag after which the ACF value is below threshold (between ±0.15) and tails off.

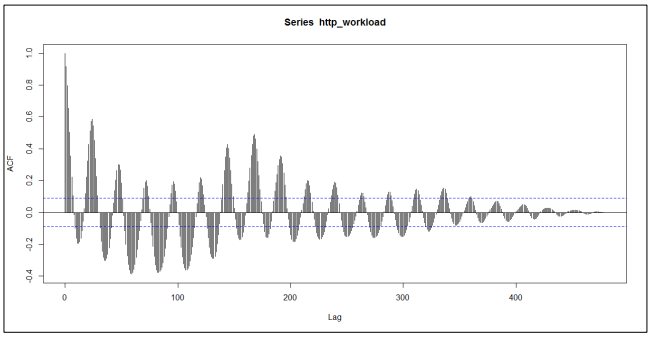

Fig. 2. Auto Co-relation Function(ACF)

The series is differenced and ACF checked at each stage and after twice differencing, the ACF and PACF tail off to zero. Figure 3 shows time series of http request after 2nd difference, which shows mean is almost constant over the lag.

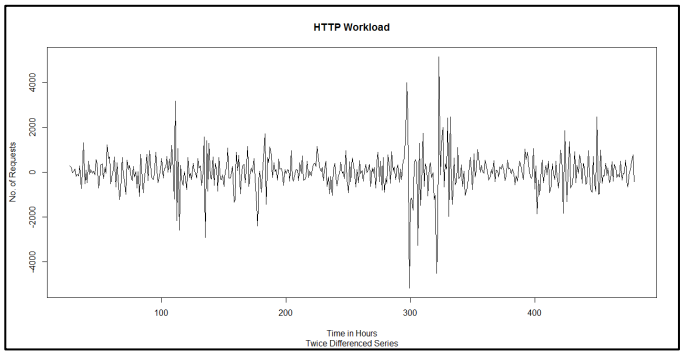

Fig. 3. Time series of http request after 2nd Difference

Model Identification

In the first step, we try to eliminate the seasonal component by differencing the series. In the next step, we find the seasonal orders P and Q. The order is chosen such that it not only captures all the information of time series but also less complex. Akaike Information criterion (AIC) [12] value gives the trade-off between model complexity and accuracy. Smaller the AIC value, more parsimonious (less complex) and better the model is. P is the order of seasonal auto-regressive (AR) model which is equal to the lag after which the PACF value is below threshold and tails off. Since the threshold value is crossed at 24, 48, 72 and 120 hours, the order of AR process (P) can be 3 to 5.

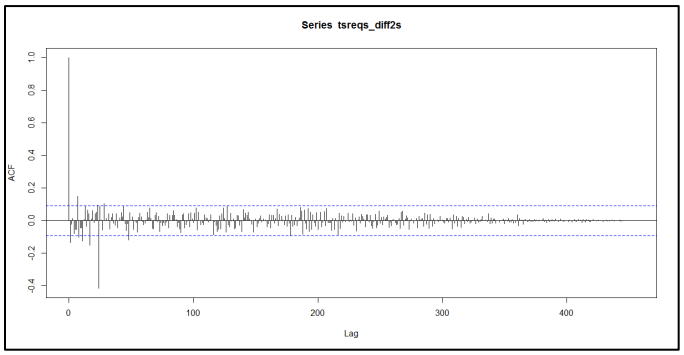

Fig. 4. ACF after second Difference

Figure 4 shows ACF of series after second difference. Which shows peak at the lag of 24 hours which shows order of MA is 1. Figure 5 shows PACF of series after second difference. Which shows peak at the lag of 24 hours which shows order of MA is 1. Similarly Q is the order of seasonal moving average (MA) model which is equal to the lag after which the ACF value is below threshold and tails off. Since the ACF plot has an overshoot only at 24 hours, the order of seasonal MA process (Q) can be 1. The order of AR (p) and MA process (q) can be computed by looking at PACF and ACF plots till lag 23. Since the values tail off, no differencing is required and hence d is 0. The AR and MA orders (p and q) may be chosen 0 or 1 based on AIC value. A table showing various seasonal and ARMA orders combinations and corresponding AIC values are shown in the table below.

Seasonal Order (P,D,Q)	ARIMA order (p,d,q)	AIC value
(3,2,1)	(0,0,0)	7286.31
(3,2,1)	(0,0,1)	6982.16
(3,2,1)	(1,0,0)	6955.64
(4,2,1)	(1,0,0)	6950.69
(5,2,1)	(1,0,0)	6938.96

Table. 1. AIC values for Model Selection

Since there is not much change in AIC value from seasonal AR order 3 to 5, order (1, 0, 0) x $(3, 2, 1)_{24}$ is chosen for forecasting.

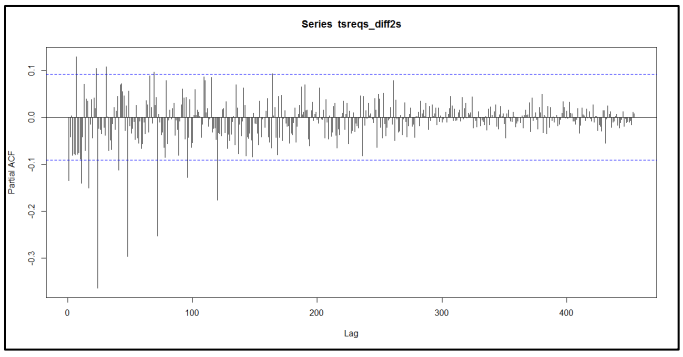

Fig. 5. PACF after second Difference

Parameter Estimation

Parameter estimation is performed by implementing maximum likelihood criterion. Given n observations, the likelihood function is defined as the probability of obtaining the data actually observed. In present work, the parameter values are computed using statistical software R. The software computes coefficients and corresponding standard error. If the absolute ratio of coefficient to error is greater than 2 then the parameter value can be accepted. The results are shown in the table:

	ar1	sar1	sar2	sar3	sma1
Coefficients	0.9066	-0.6107	-0.4534	-0.2238	-1.0000
S.E.	0.0216	0.0501	0.0536	0.0485	0.0436

Table. 2. Parameter Coefficients & corresponding Errors

The parameter estimation is as follows using equation 1.

$$(1-0.9066B)(1-B^{24})(1-B^{24})(1-(-0.6107)B^{24})$$
$$(1-(-0.4534)B^{48})(1-(-0.2238)B^{72})X_t$$
$$=(1-(-1)B^{24})w_t$$

where B is the backward operator

From the equation, future value of X_t may be forecast which is dependent on 17 terms.

Forecasting

From the equation 1, one can forecast the future data. Figure 6 shows forecast using ARMA (2, 1). Clearly ARMA model doesn't fit at all to capture the information buried in the data while seasonal ARIMA forecasts the series quite well. Figure 7 shows forecast for next 7 days using SARIMA (1, 0, 0) x (3, 2, 1)$_{24}$.

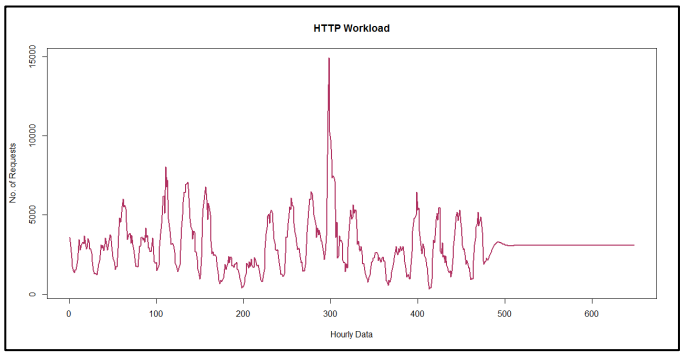

Fig. 6. Forecasting Using ARMA(2,1)

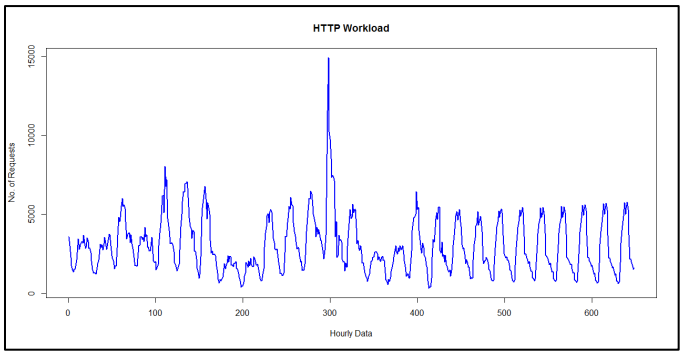

Fig. 7. Forecasting Using SARIMA (1, 0, 0) x (3, 2, 1)$_{24}$

Figure 8 shows graph of comparison of ARMA (2, 1) & SARIMA (1, 0, 0) x (3, 2, 1)$_{24}$ with Actual data. This data is collected for 20 days & next 7 days data is forecasted. In which black line shows actual data, maroon line shows forecasted data with Auto Regressive Moving Average Model & Blue line Shows forecasted data with Seasonal Auto Regressive Integrated Moving Average Model.

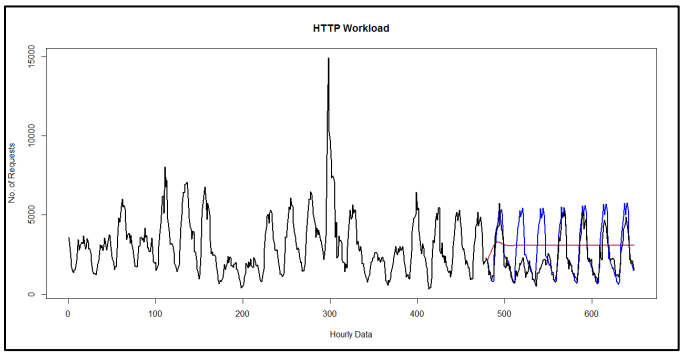

Fig. 8. Comparison of ARMA (2, 1) & SARIMA (1, 0, 0) x (3, 2, 1)$_{24}$ with Actual data

Diagnostic Check

The goodness of fitted model is checked for by testing the residuals of the model for actual workload. The independence of the residuals ensures that there is no more information in the data that can be extracted. The standardized residuals are plotted in figure. 9 and their ACF are shown in figure. 10 for the model. There is no evident pattern in the residuals which is a good sign of independence.

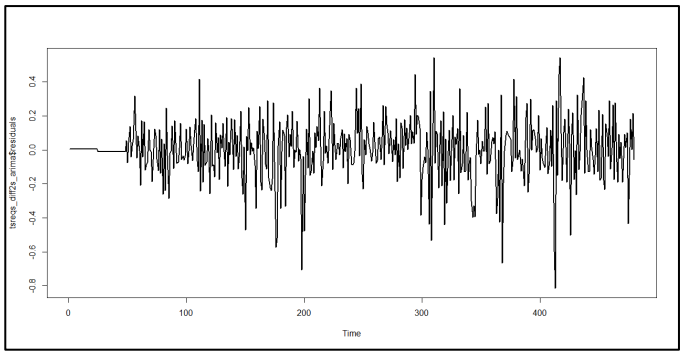

Fig. 9. Standard residuals of SARIMA

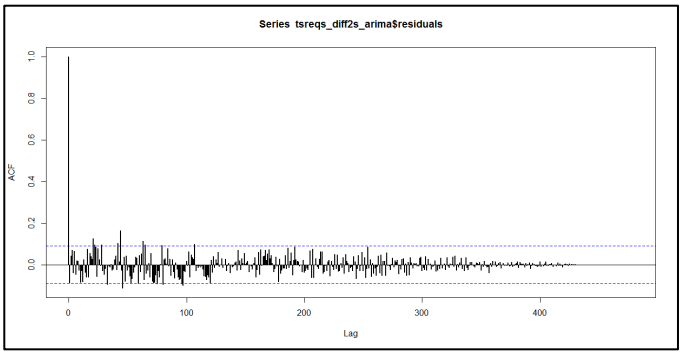

Fig. 10. ACF for standard residuals SARIMA

Since the ACF is almost zero at all lags there is no correlation among the residuals which is a good sign for their independence. If the dependence is found among residuals the process is iterated and new model is searched for fitting the given workload. Figure 11 shows ACF for residuals of ARMA, which are correlated.

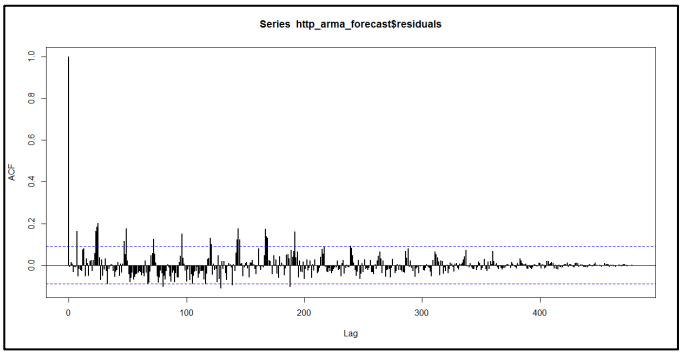

Fig. 11. ACF for standard residuals ARMA

The percentage error was computed using the metric Symmetric Mean Absolute Percentage Error (SMAPE) [16] which is given by:

$$SMAPE = \frac{\sum \left| X_t - \hat{X}_t \right|}{\sum \left| X_t + \hat{X}_t \right|} \qquad \text{Equation 2.}$$

where X_t is the actual workload and $\hat{X}_t$ is the forecasted workload.

The total error computed was 14.17% for 0% confidence interval for Seasonal Auto Regression Integrated Moving Average Model which was 24.78 % for Auto Regression Moving Average Model. Forecast data was compared with actual data. If resources are allocated according to higher confidence interval, there is a fair chance of no SLA violation.

IV. CONCLUSION

In this paper, we have used Seasonal Auto Regressive Integrated Moving Average Model for forecasting future workload. We concluded that for variable workload conditions, one needs to predict the future workloads and allocate the resources accordingly. Forecasting based policy is superior to reactive policy and will be the component of future cloud resource provisioning technologies. As if we can predict the future workload, then we can provide required resources to applications running on Infrastructure of a cloud well in advance to give better response time to cloud users.

REFERENCES

M. Armbrust, A. Fox, R. GriFFIth, A. D. Joseph, R. Katz, A. Konwinski, G. Lee,D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “A view of cloud computing,"Commun. ACM, vol. 53, no. 4, pp. 50{58, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721672

P. Mell and T. Grance, The nist definition of cloud computing," National Institute of Standards and Technology [Online]. Available:http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

S.K. Sowmya, P. Deepika, J. Naren,”Layers of Cloud – IaaS, PaaS and SaaS: A Survey”, *International Journal of Computer Science and Information Technologies*, Vol. 5 (3) , 2014, 4477-4480

Jens Myrup Pedersen, M. Tahir Riaz, Bozydar Dubalski, Damian Ledzinski, Joaquim Celestino Júnior, Ahmed Patel,”Assessing Measurements of QoS for global Cloud Computing Services” *Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing,2011,pp. 682-689*

André Pessoa Negrão, Miguel Adaixo, Luís Veiga and Paulo Ferreira, “On demand Resource Allocation Middleware for Massively Multiplayer Online Games”, IEEE 13th International Symposium on Network Computing and Applications,2014 pp. 71-74

Z. Gong, X. Gu, and J. Wilkes, “Press: Predictive elastic resource scaling for cloud systems,” in *Proceedings of the 6th Intl. Conference on Network and Service Management*, ser. CNSM 2010. IEEE, 2010, pp.9-16

N. Roy, A. Dubey, and A. Gokhale, “Efficient autoscaling in the cloud using predictive models for workload forecasting,” in *Proceedings of the 4th Intl. Conference on Cloud Computing*, ser. CLOUD 2011. IEEE, 2011, pp. 500–507.

Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd Symposium on Cloud Computing*, ser. SOCC 2011. ACM, 2011, pp. 5:1–5:14.

Ali Yadavar Nikravesh, Samuel A. Ajila, Chung-Horng Lung, "Cloud Resource Autoscaling System based on Hidden Markov Model (HMM)", IEEE International Conference on Semantic Computing, 2014 pp. 124-127

W. Fuller, Introduction to statistical time series, ser. A Wiley publication in applied statistics. New York [u.a.]: Wiley, 2013. [Online].Available:http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+021846995&sourceid=fbw bibsonomy

Y.-W. Cheung and K. S. Lai, "Lag order and critical values of the augmented dickey-fuller test," Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 277–80, July 1995. [Online]. Available: http://ideas.repec.org/a/bes/jnlbes/ v13y1995i3p277-80.html

"Akaike information criterion," 2015. [Online]. Available: http://en.wikipedia.org/ wiki/Akaike information criterion

R. J. H. with contributions from Slava Razbash and D. Schmidt, forecast:Forecasting functions for time series and linear models, 2014, r package version3.1.1. [Online]. Available: http://CRAN.R-project.org/package=forecast

F. Li and P. Luan, "Arma model for predicting the number of new outbreaks of Newcastle disease during the month," in Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on, vol. 4, 2011, pp. 660-663.

"Lag Order and Critical Values of the Augmented Dickey-Fuller Test", 2015 [Online] Available.https://ideas.repec.org/a/bes/jnlbes/v13y1995i3p277-80.html

"Symmetric mean absolute percentage error", 2015 [Online] Available : http://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error