IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. II (Jan.-Feb. 2017), PP 01-05
www.iosrjournals.org

Linear Regression Model Fitting and Implication to Self Similar Behavior Traffic Arrival Data Pattern at Web Centers

D. Mallikarjuna Reddy[1] Pushpalatha Sarla[2]
[1,2,] *Department of Engineering Mathematics, GITAM University Hyderabad Campus , Hyderabad-502329, Telangana, India.*

Abstract: *We present a simple linear regression model fit in the direction of self-similarity behavior of internet user's arrival data pattern. It has been reported that Internet traffic exhibits self-similarity. Motivated by this fact, real time internet users arrival patterns considered as traffic and the results carried out and proven that it has the self-similar nature by various Hurst index methods. The present study provides a mathematical model equation in terms linear regression as a tool to predict the arrival pattern of Internet users data at web centers. Numerical results, analysis discussed and presented here plays a significant role in improvement of the services and forecasting analysis of arrival protocols at web centers in the view of quality of service (QOS).*
Keywords*: long-range dependence, self-similarity, Poisson Process, Linear Regression, Hurst Index.*

I. Introduction

Normally traffic that is bursty on many or all time scales can be described statistically using the notion of self-similarity. Self-similarity is the property we associate with fractals; the object appears the same regardless of the scale at which it is viewed [1]. In the case of stochastic objects like time series, self similarity is used in the distributional sense when viewed at varying scales, the objects distribution remains same. Since a self-similar process has observable bursts on all time scales, it exhibits long-range dependence values at any instant are typically correlated with values at all future instants; Surprisingly the self-similarity of Ethernet network traffic has been thoroughly established. The importance of long-range dependence in network traffic is beginning to be observed in studies such as [2], [5] which show that packet loss and delay behavior is radically different in simulations using real traffic data rather than traditional network models. While the understanding the nature of network traffic is significant in order to know the design and implement computer networks and network services like the World Wide Web (WWW) [3]. Recent examinations of LAN traffic and wide area network (WAN) traffic have challenged the commonly assumed models for network traffic [4], *e.g.*, the Poisson process. Were traffic to follow a Poisson or Markovian arrival process (MAP), it would have a characteristic burst length which would tend to be smoothed by averaging over a long enough time scale. Rather, measurements of real traffic indicate that significant traffic variance (burstiness) is present on a wide range of time scales [6]. Recently Poison process has been used to model the highway traffic irrespective of traffic intensity [7-11].

The scheme of this paper is first, we provide a mathematical model equation fit in terms linear regression [16] as a tool to predict the arrival pattern of Internet users data at web centers, second, using the fitted data we examine whether web users traffic data has again the self similar behavior. This is an enhancement of prior results carried out using a real time series data [12-15]. This kind of analysis is often useful in improvement of the services and forecasting analysis of arrival pattern at various organizations. The rest of the paper is organized as follows: Definition of self-similarity or long range dependence is given in Section II. Statistical Tests for Self-Similar Characteristics: placed in Section III. In Section IV, Linear Regression Model Fitting and Analysis of data discussed. Finally, conclusions are given in Section V.

II. Definitions Of Self-Similarity And Long-Range Dependence

In this section we give brief idea of the mathematical basis for self-similar processes and long-range dependence.

Self-Similarity: Self-similarity, in general, describes a phenomenon where a certain property of an object is preserved with respect to scaling in space or time. Let $X = \{X_t \; t = 1, 2, 3, ...\}$ be a discrete time covariance stationary stochastic process or a stationary time series with mean μ and variance σ^2.In terms of arrival pattern, X_t can be described as number of users logged on to an Internet server each minute overtime period t on a day. Given the stochastic process X, we can define the m -aggregated stochastic process

$X^m = \{X_i^m \; i = 1,2,3,...\}$ where m is a positive integer, by averaging the original stochastic process X over non overlapping blocks of size m , namely

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \; t = 1, 2,... \tag{1}$$

is the average of the original sequence in m non-overlapping blocks. Then the process X is defined as self-similar process with the Hurst parameter,

$H = 1 - \beta/2$ if $Var(X^{(m)}) = \sigma^2 m^{-\beta}, \forall m \geq 1.$ (2)

Exact second-order self-similar process:

The exact second-order self-similar process is defined as follows. Arrival instants are modeled as point process. Divide the time axis into disjoint intervals of unit length and let $X = \{X_t : t = 1,2,....\}$ be the number of points (arrival) in the t^{th} interval. Let X be a second order stationary process with variance σ^2 and the autocorrelation function $\gamma(k), \; k \geq 0$ is given by

$$\gamma(k) = \frac{Cov(X_t, X_{t+k})}{Var(X_t)} \tag{3}$$

For each $m = 1,2,3,....,$ let a new time series $X_t^{(m)}$ is obtained averaging the original time series X over non-overlapping blocks of size m. That is

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \; t = 1,2,... \tag{4}$$

This new series $X_t^{(m)}$, for each m, is also a second order stationary process with autocorrelation function $\gamma^{(m)}(k)$.

Definition (i): The process $'X'$ is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\gamma(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \tag{5}$$

Definition (ii): The process $'X'$ is said to be asymptotically second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\sum_{m\to\infty} \gamma^{(m)}(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \tag{6}$$

In terms of variance, self- similar process is defined as follows:

Definition (iii): The process $'X'$ is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$Var(X^{(m)}) = \sigma^2 m^{-\beta}, \forall m \geq 1 \tag{7}$$

Now we shall differentiate long range dependence (LRD) and short range dependence (SRD) processes. For $H \neq 0.5,$ from the Eq.(5), we can see that $\gamma(k) = H(2H-1)k^{2H-2}$ as $k \to \infty$, and we have

$$\sum_k \gamma(k) \sim c\sum_k k^{-\beta}, \; c = H(2H-1).$$

(8) The series $c\sum_k k^{-\beta}$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$ otherwise they are convergent, being a positive term series. Accordingly the left hand series $\sum_k \gamma(k)$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1,$ otherwise they are convergent. That is, for $0.5 < H < 1$, the autocorrelation functions decays slowly, that is

hyperbolically. In this case, the process X is called Long Range Dependent (LRD). The process X is Short Range Dependent (SRD) if $0 < H < 0.5$ and the autocorrelation function is summable (finite).

III. Tests For Self Similar Variable Traffic:

In assessing the self similarity for the number of users logged on to an Internet server each minute over 100-minutes time series data four statistical methods have been applied to measure Hurst index which gives the intensity of self-similarity [15]. The parameter H has range $0.5 \le H \le 1$. Using Periodogram Analysis computed value of H is 0.763. Obtained value of index H by Correlogram method is 0.79. Whereas in this case of Percentile method estimated value of H is 0.762 and by means of Variance Time analysis obtained value of H is 0.76 [16].

IV. Linear Regression Model Fitting And Analysis Of Data:

In this section, we present some practical ideas and analysis of linear regression model fitting for self-similar variable traffic arrival data pattern at various web centers. Some applications of regression involve regressor and response variables that have a natural sequential order over time. Such data are called time series data. Regression models using time series data occur relatively often in economics, business, and some fields of engineering [17-18]. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable/outcome variable. The variable we are using to predict the other variable's value is called the independent variable or the predictor variable. For instance, we could use linear regression curves to understand whether test performance can be predicted based on review time; Here, how the data exhibit a significant degree of scatter and the strategy is to derive a single curve that represents the general trend of the data? The linear regression model equation and procedures, a complete explanation of the output by this model is explained.

In lines of statistical modeling the linear regression model equation of time series data is

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{9}$$

β_0 and β_1 are model parameters.

Where $\varepsilon = Y - \hat{Y}$, ε is the error or noise term, Y and $\hat{Y}$ are respectively, the observed and predicted values of the response variable for each individual of data set. Error terms often assumed independent observations from a $N(0, \sigma^2)$ distribution. Clearly, by assumptions in linear regression are that the error terms ε has mean zero i.e. $E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2$ and uncorrelated.

Thus $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ and $E(Y) = \beta_0 + \beta_1 x$ (10)

For the time series data, we will estimate the model parameters with coefficients

$$E(Y / X) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{11}$$

Parameters are fit to minimize the sums of squared errors: that is

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \text{ is minimum.} \tag{12}$$

The coefficients β_0 and β_1 that minimize E must satisfy the following conditions:

$$\frac{\partial E}{\partial \beta_0} = 0, \quad \frac{\partial E}{\partial \beta_1} = 0$$

(13) on simplification we obtain are model parameters β_0 and β_1.

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \text{ and } \beta_0 = \bar{Y} - \beta_1 \bar{X}\ . \tag{14}$$

Thus the model parameters are estimated in liner regression curve using Eqn.8-14.

Below by using SPSS tool, we focus on the results for the linear regression analysis of self- similar variable traffic arrival data pattern at various web centers only. The model Summary and analysis of results discussed, as shown in tables. Table 4.1 provides the R and R^2 values. The R value is 0.669, which indicates a high degree of correlation of data. The R^2 value total variation in the dependent variable i.e. internet users, can be explained by the independent variable, say time. In this case, 44.7% can be explained. Also most of the

regression problems involving time series data exhibit positive autocorrelation, the hypotheses usually tested by Durbin-Watson test (d=0.028) and conclude that the errors are positively auto correlated. Analysis of variance (ANOVA), which reports well about the regression model equation fits the data (i.e., predicts the dependent variable) is shown in table 4.2. Also it reveals, the regression model predicts the dependent variable significantly well. Since the statistical significance of the regression model that was run. Here, i.e. F(1, 98) = 24.754, p = .0001, which is less than 0.05, and indicates that, overall, the regression model statistically significant and predicts the outcome variable and it is a good fit for the data. The Coefficients table-4.3 provides the information to predict the dependent variable with respect to independent variable time, as well as determine whether time in minutes contributes statistically significance to the model since observed *t-statistic* is 4.975 and *p value* < 0.01. Residual statistics for dependent variable (Internet users) shown in table-4.4.

A linear regression model equation established that time in minutes to predict number of users log on to internet server. The real time series data exhibited a significant degree of scatter plot and the linear regression model curve that represents the general trend of the data is depicted in Figure. 4.1 and the predicted regression model fit shown in Figure 4.2.

$$Y = 105.9 + 0.616X \tag{15}$$

Table-4.1

Model	*R*	R^2	*Adj* R^2	*Std. Error of the Estimate*	*Change Statistics*					*Durbin-Watson*
					R^2	*F*	*df1*	*df2*	*Sig.*	
1	.669[a]	.447	0.431	35.769	.202	24.754	1	98	.000*	.028*

Table-4.2

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1	Regression	31670.923	1	31670.923	24.754	.000*
	Residual	125384.437	98	1279.433		
	Total	157055.360	99			

Table-4.3 Dependent Variable: Internet_Users, * *Significant at 0.05 levels.*

Model		*Un standardized Coefficients*		*Standardized Coefficients*	*t*	*Sig.*	*95.0% Confidence Interval for B*	
		B	*Std. Error*	*Beta*			*Lower Bound*	*Upper Bound*
1	(Constant)	105.946	7.208		14.699	0.000	91.642	120.250
	Time in_min	0.616	0.124	0.449	4.975	0.000*	.371	.862

a. Dependent Variable: Internet_Users

Table-4.4

Residual Statistics	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*	*N*
Predicted Value	106.56	167.60	137.08	17.886	100
Residual	65.952	62.252	.000	35.588	100
Std. Predicted Value	1.706	1.706	.000	1.000	100
Std. Residual	1.740	1.844	.000	.995	100

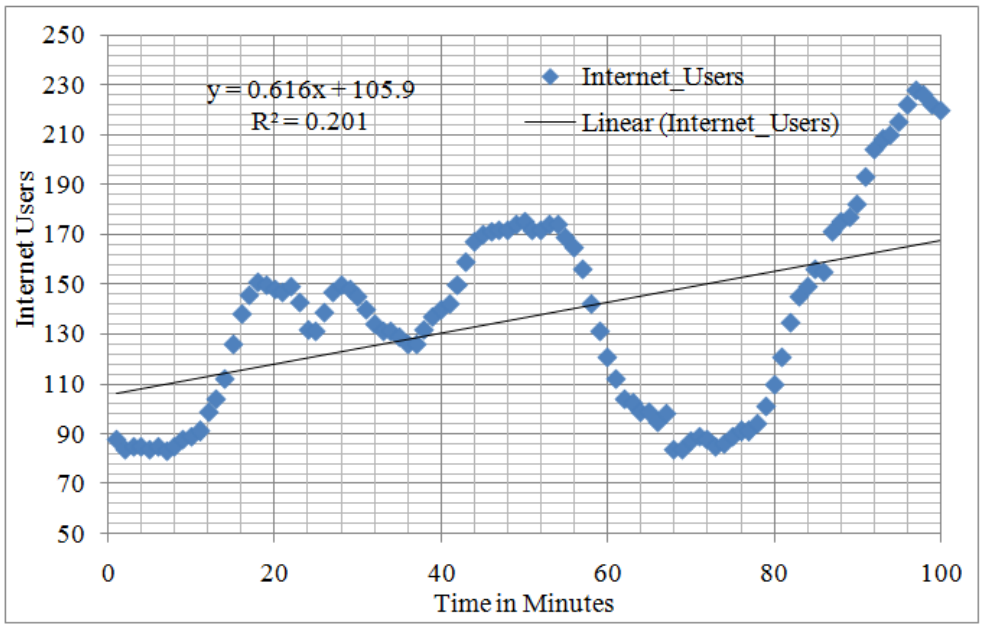

Figure 1 Time (in minutes) Vs Internet Users

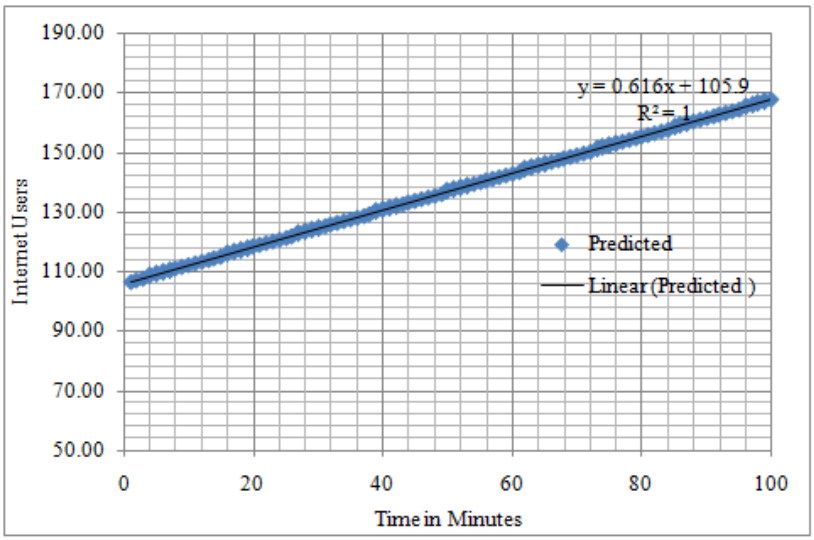

Figure 2: Time (in minutes) Vs Internet Users

V. Conclusions

In this paper, self-similar behavior time series data has been considered as traffic data set from various web centers. A mathematical model equation fitted in terms linear regression as a tool to predict the arrival pattern of Internet user's data at web centers. These kind of fitting models offered here can plays significant role in improvement of the services and forecasting analysis of arrival patterns at web centers in the view of quality of service (QOS).

References

[1]. W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson On the Self-Similar Nature of Ethernet Traffic (Extended version) , IEEE / ACM Trans. Networking, 2, pp. 1-15, 1994.
[2]. J.Beran, M. S. Taqqu and W. Willinger, Long- range dependence in variable bit rate traffic," IEEE Trans. on Communications, Vol. 43, pp.1566-1579
[3]. M. Crovella and A.Bestavros, Self-Similarity in World Wide Web traffic: evidence and possible causes, IEEE/ ACM Trans. Networking, pp.835-846, 1997.
[4]. Vern Paxson, Sally Floyd, Wide Area Traffic: The Failure of Poisson Modeling, IEEE/ACM Trans. Networking, 3, pp.226-244, 1995.
[5]. Allan T. Anderson, Bo Friis Nielsen , A Markovian Approach for Modeling Packet Traffic with Long Range Dependence, IEEE Journal on Selected Areas in Communications, Vol.16, No.5, pp.719-732, June 1998.
[6]. T. Yoshihara, S.Kasahara, and Y. Takahashi, Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process, Telecommun. Syst., vol.17, pp.185-211, 2001.
[7]. S.K. Shao, Malla Reddy Perati, M.G. Tsai, H.W. Tsao and J. Wu, Generalized variance-based Markovian fitting for self-similar traffic modeling, IEICE Trans. Commun., Vol.E88-B, no.12, pp.4659-4663, April 2005.
[8]. M. M. A. Sarker, Estimation of the Self-similarity parameter in long memory processes, Journal of Mechanical Engineering, Vol. ME38, Dec. 2007, Transaction of the Mech. Eng. Div., The Institution of Engineers, Bangladesh.
[9]. Nagatani, T. (2005). Self-similar behavior of a single vehicle through periodic traffic lights, Physica A, 347, 673–682.
[10]. Qiang Meng and Hooi Ling Khoo, Self-similar characteristics of vehicle arrival pattern on Highways. Journal of Transportation Engineering, © ASCE / November 2009.
[11]. K. Raghavendra, Malla Reddy Perati; H. K. Reddy Koppula; Mallikarjuna Reddy Doodipala, Rajaiah Dasari 'Self-Similar Behavior of Highway Road Traffic and Performance Analysis at Toll Plazas" 1234 Journal of Transportation Engineering © Asce October 2012.
[12]. D.Mallikarjuna Reddy "Second Order Statistics of Time Series of Various Real Time Problems in Conjunction with Periodogram Technique" International Journal of Latest Trends in Engineering and Technology (IJLTET) Vol. 3 Issue 1 September 2013
[13]. Pushpalatha Sarla, D.Mallikarjuna Reddy, Manohar Dingari, "A Study on Self Similarity Analysis of Web Users Data at Selected Web Centers" Proceedings of International conference on Mathematics ICM- 2015.
[14]. P. J. Brockwell and R. A. Davis, \An introduction to time series and fore-casting", Springer - Verlag, New York (1996).
[15]. Sarla, P. , Doodipala, M. and Dingari, M. (2016) Self Similarity Analysis of Web Users Arrival Pattern at Selected Web Centers. American Journal of Computational Mathematics, 6, 17-22. doi: 10.4236/ajcm.2016.61002
[16]. Pushpalatha Sarla, D.Mallikarjuna Reddy, Manohar Dingari " Queue Length-Busy Time Distribution Of Web Users Data With Self Similar Behavior Volume: 05 Special Issue: 05 | ICIAC-2016 | May-2016,
[17]. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). Introduction to Linear regression Analysis. 3rd Edition, New York, New York: John Wiley & Sons.
[18]. Spyros Markidakis, steven C.Wheelwright, Rob J.Hydman "Forcating Methods and Applications" John Wiley & Sons.Inc. Third edition".