Forecasting Internet Traffic by Using Seasonal GARCH Models

Sahm Kim

Abstract: **With the rapid growth of internet traffic, accurate and reliable prediction of internet traffic has been a key issue in network management and planning. This paper proposes an autoregressive-generalized autoregressive conditional heteroscedasticity (AR-GARCH) error model for forecasting internet traffic and evaluates its performance by comparing it with seasonal autoregressive integrated moving average (ARIMA) models in terms of root mean square error (RMSE) criterion. The results indicated that the seasonal AR-GARCH models outperformed the seasonal ARIMA models in terms of forecasting accuracy with respect to the RMSE criterion.**

Index Terms: **Akaike information criterion (AIC), Internet traffic, root mean square error (RMSE), seasonal autoregressive-generalized autoregressive conditional heteroscedasticity (AR-GARCH), seasonal autoregressive integrated moving average (ARIMA).**

I. INTRODUCTION

Statistical time series models have been very effective tools for forecasting finance- and business-related data. With the rapid growth of internet traffic, the accurate and reliable prediction of internet traffic data has been a key issue in network management and planning.

Prediction methods including statistical tools have played major roles in analyzing and forecasting internet traffic. Autoregressive integrated moving average (ARIMA) models, proposed by Box *et al.* [1], have been used for forecasting time series data in many fields. Recently, internet traffic has been analyzed and predicted using various time series models, including ARIMA models.

The main characteristics of internet traffic include long-memory properties and nonstationary ones such as seasonality and heavy-tailed distributions of errors in models. Jiakun *et al.* [2] applied a factional ARIMA (FARIMA) model to explain the long-memory property of network traffic. On the other hand, Shu *et al.* [3] used seasonal ARIMA (SARIMA) models to forecast mobile traffic in China. Krithikaivasan *et al.* [4] used autoregressive conditional heteroscedasticity (ARCH) models to forecast periodically nonstationary traffic and for dynamic bandwidth provisioning.

In this paper, we propose seasonal generalized ARCH (SGARCH) models to forecast internet traffic and demonstrate that the proposed models outperform SARIMA models in terms of the root mean square error (RMSE) criterion. The rest of this paper is organized as follows: Section II introduces statistical forecasting models, namely SARIMA models, and seasonal autoregressive-generalized autoregressive conditional heteroscedasticity (AR-GARCH) models. Section III presents the results of the performance evaluation of the proposed models in terms of the RMSE criterion, and Section IV concludes.

Manuscript received May 9, 2011.

This Research was supported by the Chung-Ang University Research Grants in 2009.

The author is with the Department of Applied Statistics, Chung-Ang University, 221 Heukseok-Dong, Dongjak-Gu, Seoul 156-756, Korea, email: sahm@cau.ac.kr.

II. STATISTICAL TIME SERIES MODELS

Many time series models have been employed to analyze and predict various types of data in the fields of finance, economics, biology, and engineering. To analyze and predict internet traffic accurately and reliably, we first introduce the ARCH model originally proposed by Engle [5] to explain the volatility mainly in financial time series data.

Recently, some researchers such as Zhou *et al.* [6] have demonstrated that the volatility patterns of internet traffic are quite similar to those of financial time series data. This indicates the usefulness of ARCH models for analyzing and forecasting internet traffic. In the next section, we first define SARIMA and AR-GARCH models and then present SGARCH models.

A. Seasonal ARIMA Models

One of the most popular and useful methods for analyzing and predicting internet traffic is the ARIMA model [7]. First, we define the ARIMA (p, d, q) model as

$$\phi(B)\nabla^d Z_t = \theta(B)e_t$$

where Z_t is the observed traffic at time t, $\nabla^d Z_t = (1-B)^d Z_t$, $B^j Z_t = Z_{t-j}$, and e_t is an independently and identically distributed random variable with mean 0 and constant variance; d is a non-negative integer; B is a backward shift operator; and ∇ is a differencing operator. In the ARIMA model, the seasonal differencing is often used to treat seasonal patterns in data. Suppose that the seasonal period is s. Then, the seasonal differencing is defined as

$$\nabla_s Z_t = Z_t - Z_{t-s}$$

where ∇_s is the differencing operator for period s. By this procedure, there is no seasonality for the series $\{\nabla_s Z_t : t = s+1, \cdots, n\}$. The pure seasonal ARIMA model is defined as

$$\begin{aligned}
&\Phi(B^s)\nabla_s^D Z_t = \Theta(B^s)a_t \\
&\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \\
&\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}. \quad (1)
\end{aligned}$$

In general, D is 1, and a_t is assumed to be a seasonally adjusted series that can be represented by a nonseasonal

1229-2370/11/$10.00 © 2011 KICS

ARIMA (p, d, q) model, that is,

$$\begin{aligned}\phi(B^s)\nabla^d a_t &= \theta(B)b_t \\ \phi(B^s) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\ \theta(B^s) &= 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \end{aligned} \quad (2)$$

and b_t is assumed to be an independent and identically distributed normal random variable with mean 0 and variance σ_b^2.

By combining (1) with (2), we have the following seasonal ARIMA $(p, d, q) \times (P_{s_1}, D_{s_1}, Q_{s_1})_{s_1} \times (P_{s_2}, D_{s_2}, Q_{s_2})_{s_2}$ model:

$$\phi(B)\Phi(B^{s_1})\Pi(B^{s_2})\nabla_{s_1}^{D_1}\nabla_{s_2}^{D_2}\nabla^d Z_t = \theta(B)\Theta(B^{s_1})\Psi(B^{s_2})b_t.$$

The parameters of the models are typically estimated by the maximum likelihood estimation method, and the optimal number of parameters is determined by the Akaike information criterion (AIC) [8].

B. AR-GARCH Models

Bollerslev [9] proposed the generalized ARCH (GARCH) model, whose main feature is that it can be fitted to data with heavier-tailed error distributions. The AR (k)-GARCH (p, q) model is defined as

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \epsilon_t \\ \epsilon_t &= \sqrt{h_t} e_t \\ h_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \end{aligned} \quad (3)$$

where $e_t \sim$ i.i.d. $N(0, \sigma^2)$, $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\alpha_1 + \beta_1 < 1$ such that the model is weakly stationary. Zhou *et al.* (2005) showed that the ARIMA-GARCH model exhibits better prediction accuracy than the FARIMA model for forecasting network traffic.

C. Seasonal AR-GARCH Model

Seasonal AR-GARCH models are given by the following equations.

$$\begin{aligned} \phi(B)\Phi_{p_1}(B^{s_1})\Pi_{p_2}(B^{s_2})Z_t &= \epsilon_t, \epsilon_t \\ &= e_t\sqrt{h_t}, e_t \sim \text{i.i.d. N}(0, \sigma^2) \\ h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 &+ \sum_{j=1}^{q} \beta_j h_{t-j} + \sum_{i=1}^{p_1} \alpha_{is_1} \epsilon_{t-is_1}^2 \\ + \sum_{j=1}^{q_1} \beta_{js_1} h_{t-js_1} &+ \sum_{i=1}^{p_2} \alpha_{is_2} \epsilon_{t-is_2}^2 + \sum_{j=1}^{q_2} \beta_{js_2} h_{t-js_2} \end{aligned} \quad (4)$$

where s_1, s_2 refers to the stage of the seasonal cycle at time t. The main reason for considering the model is that we need to identify seasonal patterns of internet traffic. The equations in (4) are quite complicated. However, in the next section we will represent the seasonal AR-GARCH model precisely and evaluate its performance in comparison with the seasonal ARIMA models.

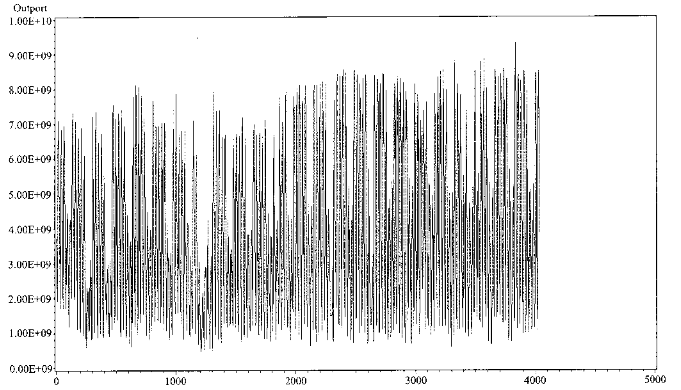


Fig. 1. Time plot for original data.

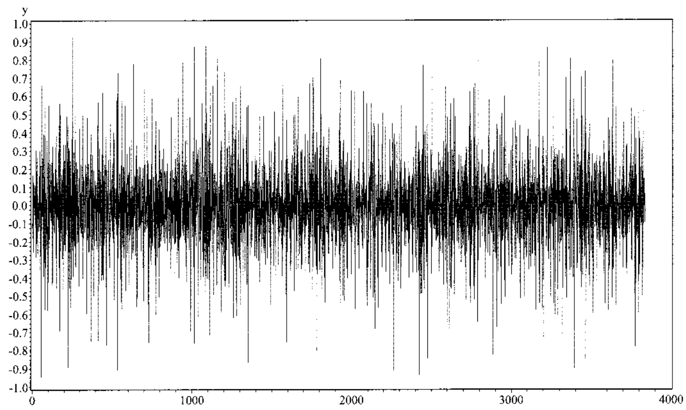


Fig. 2. Time plot for log-differenced data.

III. PERFORMANCE EVALUATION AND DISCUSSION

Data is collected from the link connecting Chung-Ang University (Seoul, Korea) to outside world between December 15, 2010 and June 2, 2011. The data is measured every 5 minutes and 12 measurements per hour are aggregated to yield a single data point per hour. This results in 4080 data points for the 170 day window. In Figs. 1 and 2, we present the original data set and the log-differenced data set which is defined by $z_t = \log(x_t) - \log(x_{t-1})$, where x_t and x_{t-1} are the observed traffic at time t and time $t - 1$, respectively. The main reason for transforming the original data is to guarantee the stationarity of the data to fit the time series models. For the data sets, we have 4080 hourly data points; of these, 3744 data points (for 156 days) were used to build up the models and 336 data points (for two weeks) for comparing the performance evaluation of the models. We can easily see the seasonal patterns for the data sets and know that the periods are 24 and 168, which correspond to daily and weekly cycles, respectively.

In Table 1, we have results of the Lagrange multipliers (LM) test, which was based on Breusch and Pagan [10], they show strong evidence of the heteroscedastic variance in the data.

Table 2 presents results for various SARIMA models and Seasonal AR-GARCH models in terms of their performance and list the AIC values for the optimal models. Table 3 shows the parameter estimates and RMSE values. Based on the results in

Table 1. LM test.

Order	LM	p value
1	71.2031	<0.0001
2	90.9608	<0.0001
3	111.4075	<0.0001
4	112.4240	<0.0001
5	112.4336	<0.0001
6	114.6980	<0.0001
7	118.0427	<0.0001
8	123.1657	<0.0001
9	125.1147	<0.0001
10	132.0991	<0.0001
11	137.9747	<0.0001
12	140.1722	<0.0001

Table 2. Fitted models.

Model	AIC
SARIMA $(1,0,2)(1,0,1)_{24}(0,0,1)_{168}$	−4639.12
SARIMA $(1,0,2)(0,0,1)_{24}(0,0,1)_{168}$	−4639.21
SARIMA $(2,0,3)(1,0,1)_{24}(0,0,1)_{168}$	−4646.96
AR $(3)(1)_{24}(1)_{168}$ GARCH $(1,1)(1,1)_{24}(0,0)_{168}$	−2725.39
AR $(3)(1)_{24}(1)_{168}$ GARCH $(0,0)(1,1)_{24}(0,0)_{168}$	−3570.23
AR $(3)(1)_{24}(1)_{168}$ GARCH $(1,0)(1,1)_{24}(0,0)_{168}$	−3607.10
AR $(3)(2)_{24}(1)_{168}$ GARCH $(0,1)(1,1)_{24}(0,0)_{168}$	−3796.48
AR $(3)(2)_{24}(2)_{168}$ GARCH $(0,1)(1,1)_{24}(0,0)_{168}$	−3916.43
AR $(2)(2)_{24}(2)_{168}$ GARCH $(0,1)(1,1)_{24}(0,0)_{168}$	−3896.37
AR $(3)(2)_{24}(2)_{168}$ GARCH $(1,1)(0,0)_{24}(0,0)_{168}$	−3408.56
AR $(3)(2)_{24}(2)_{168}$ GARCH $(0,1)(0,0)_{24}(1,1)_{168}$	−3868.84
AR $(3)(2)_{24}(2)_{168}$ GARCH $(1,1)(1,1)_{24}(1,1)_{168}$	−3457.66

Table 2, SARIMA $(2,0,3)(1,0,1)_{24}(0,0,1)_{168}$ was the best-performing SARIMA model in terms of AIC values:

$$(1-\phi_1 B-\phi_2 B^2)(1-\Phi_{24}B^{24})y_t = (1-\theta_1 B-\theta_2 B^2-\theta_3 B^3)(1-\Theta_{24}B^{24}) \cdot (1-\Psi_{168}B^{168})\epsilon_t. \quad (5)$$

The second column of Table 3 presents the parameter estimates for the model. We also examined the seasonal AR-GARCH model to capture the seasonal patterns in the heteroscedastic conditional variance; the optimal AR-GARCH model is of the form

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-\Phi_{24}B^{24}-\phi_{48}B^{48}) (1-\Pi_{168}B^{168}-\pi_{336}B^{336})y_t=\epsilon_t$$

where $\epsilon_t=\sqrt{h_t}e_t$.

$$h_t=\alpha_0+\beta_1 h_{t-1}+\alpha_{24}\epsilon_{t-24}^2+\beta_{24}h_{t-24}, \quad \epsilon_t \sim \text{i.i.d. } N(0,\sigma^2) \quad (6)$$

and the parameter estimates for the model are shown in the Table 3.

We compared the prediction accuracy of the models based on the RMSE, which can be defined as follows:

Table 3. Parameter estimates and RMSE.

Parameter	Estimate	S.E.	p value
Seasonal ARIMA			
ϕ_1	−0.2615	0.0199	<0.0001
ϕ_2	0.7179	0.0175	<0.0001
Φ_{24}	−0.0277	0.0209	0.1852
θ_1	−0.0958	0.0258	0.0002
θ_2	0.9664	0.0143	<0.0001
θ_3	0.0680	0.0225	0.0025
Θ_{24}	0.8062	0.0128	<0.0001
Ψ_{168}	0.7800	0.0108	<0.0001
RMSE		0.3727	
Seasonal AR-GARCH			
ϕ_1	0.0686	0.0107	<0.0001
ϕ_2	0.0406	0.0103	<0.0001
ϕ_3	0.1017	0.0094	<0.0001
Φ_{24}	0.3928	0.0158	<0.0001
Φ_{48}	0.1281	0.0151	<0.0001
Π_{168}	0.4337	0.0145	<0.0001
Π_{336}	0.2819	0.0132	<0.0001
α_0	0.0028	0.0003	<0.0001
α_{24}	0.3806	0.0111	<0.0001
β_1	0.2404	0.0062	<0.0001
β_{24}	0.2658	0.0123	<0.0001
RMSE		0.2098	
RMSE ratio		0.2098/0.3727 = 0.5628	

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\widehat{y}_t)^2} \quad (7)$$

where y_t is the real value of data at time t; $\widehat{y}_t$ is the predicted value of data at time t; and n is the total number of estimated forecast values. For our data set, n is 336. The measure RMSE is one of the measures for comparing prediction errors for the last 336 data points.

The last row of Table 3 shows the RMSE ratio. The ratio is 0.5628. This ratio indicates that the seasonal AR-GARCH model well explained the volatility of the data set and provides evidence of the high volatility of internet traffic such as finance and business data which is consistent with the findings of many studies.

IV. CONCLUDING REMARKS

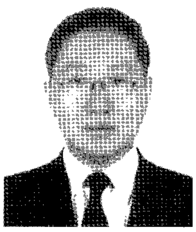
In this paper, we introduced seasonal ARIMA and AR-GARCH models and evaluated their performance in terms of predicting internet traffic at the small area. The results indicate that the seasonal AR-GARCH models outperformed the SARIMA models in terms of forecasting accuracy with respect to the RMSE criterion. In this regard, future research should consider more sophisticated GARCH models to predict internet traffic and for dynamic bandwidth provisioning. We can also consider the detection methods of the abnormal internet traffic based on the models that we have used in this paper. It may be noted that we consider one data point for each hour over 156 days to predict two weeks of traffic. Our model is, thus, useful at this granularity. However, our model is not applicable for how traffic variation within an hour could be predicted; for traffic data at finer granularity, other applicable models may be developed. Finally, we note that our model uses RMSE; for network traffic, it has been pointed out [11] that a generalized-cost function approach that penalizes under-forecasting is an important consideration in predicting network traffic. Thus, this is another direction for future work.

ACKNOWLEDGMENTS

The author is grateful for two anonymous referees for careful comments. I am also grateful for professor Medhi for his helpful and detailed comments. I also thank Myungho Ha (Chung Ang University, Seoul, Korea) for his assistance on handling the data.

REFERENCES

[1] G. E. P. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control.* Prentice Hall, 1994.
[2] L. Jiakun, S. Yantai, Z. Lianfang, and X. Fei, "Traffic modeling based on FARIMA models," in *Proc. CCECE*, vol. 1, 1999, pp. 162–167.
[3] Y. Shu, M. Yu, O. Yang, J. Liu, and H. Feng, "Wireless traffic modeling and prediction using seasonal ARIMA models," *IEICE Trans. Commun.*, pp. 1675–1679, 2005.
[4] B. Krithikaivasan, Y. Zeng, K. Deka, and D. Medhi, "ARCH-based traffic forecasting and dynamic bandwidth provisioning for periodically measured nonstationary traffic," *IEEE/ACM Trans. Netw.*, vol. 15, pp. 683–696, June 2007.
[5] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation," *Econometrica*, vol. 50, pp. 987–1007, 1982.
[6] B. Zhou, D. He, Z. Sun, and W. H. Ng, "Network traffic modeling and prediction with ARIMA/GARCH," in *Proc. HET-NETS*, 2005, p. 8.
[7] Y. Shu, Z. Jin, L. Zhang, and L. Wang, "Traffic prediction using FARIMA models," in *Proc. IEEE ICC*, 1999, pp. 891–895.
[8] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, pp. 716–723, 1974.
[9] T. Bollerlev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, pp. 307–327, 1986.
[10] T. S. Breusch and A. R. Pagan, "The Lagrange multiplier test and its applications to model specification in econometrics," *Review of Economic Studies*, vol. 47, pp. 239–254, 1980.
[11] B. Krithikaivasan, Y. Zeng, and D. Medhi, "Generalized cost function based forecasting for periodically measured nonstationary traffic," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, pp. 1105–1117, Sept. 2008.
[12] D. Tikunov, *Software, Telecommunications, and Computer Networks.* SoftCOM, pp. 1–5, 2007.
[13] J. Caiado, "Performance of combined double seasonal univariate time series models for forecasting water demand," *J. Hydrologic Eng.*, vol. 15, pp. 215–222, 2010.
[14] K. Papagiannaki, N. Taft, Z. Zhang, and C. Diot, "Long-term forecasting of Internet backbone traffic: Observations and initial models," in *Proc. IEEE INFOCOM*, 1999, pp. 1178–1188.



Sahm Kim is a Professor in the Department of Applied Statistics at Chung-Ang University, Seoul, Korea. He received his B.S. and M.S. degrees in Statistics and Computer Science from Seoul National University, Seoul, Korea, in 1983 and 1985, respectively. And he received Ph.D. degree in Statistics from the University of Georgia, Athens, Georgia, USA, in 1998. He had worked as a Research Staff at the Traffic Engineering Lab in Korea Telecom. His research interests include forecasting internet traffic and outlier detection in time series data.