Forecasting Seasonal Traffic Flows

Lionel Fillatre   Dimitry Marakov   Sandrine Vaton
Computer Science Department
ENST Bretagne, Brest, France
Email: firstname.name@enst-bretagne.fr

Abstract

The problem of seasonal traffic flow forecasting is addressed in this paper. It is shown that SARIMA time series models are particularly relevant to model a seasonal traffic flow. The SARIMA process is represented in linear state-space form and classical Kalman recursions provide on-line forecasting values. Experiments on a real traffic flow validate the method by supplying accurate forecasts.

I. Introduction

Network monitoring and diagnosis are key elements to improving network performance. In particular, forecasting network traffic is essential for network dimensioning, load balancing and traffic engineering tasks. Given a traffic flow, it is desirable to propose an on-line algorithm which provides accurate forecasting values. A natural framework for this problem involves state-space models. This promising approach requires modeling traffic flow evolution and, surprisingly, classical approaches using times series theory have been studied very little. This work proposes to investigate the relevance of seasonal time series to model the traffic flow by assuming it is composed of a trend and a seasonal pattern with short time correlations.

II. Problem statement

Seasonal AutoRegressive Integrated Moving Average (SARIMA) processes have been introduced in the literature to model time series with trends, seasonal pattern and short time correlations. Let us denote y_t the number of bytes passing through the observed link during the time interval $[(t-1)\Delta\,;\,t\Delta]$ of duration $\Delta > 0$ for $t = 0, \pm 1, \dots$ Let B be the backshift operator, whose effect on a time series y_t can be summarized as $(B^d y)_t = y_{t-d}$ for all integers d. As defined in [1], [2], a SARIMA $(p,d,q)\times(P,D,Q)_s$ process y_t verifies the equation:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \theta(B)\Theta(B^s)e_t$$

where e_t is a white noise sequence. Here, ϕ, Φ, θ and Θ are polynomial functions of degrees p, P, q and Q respectively. The term $(1-B)^d$ is used to eliminate polynomial trends and $(1-B^s)^D$ is used to eliminate seasonal patterns with the period s. The multiplicative polynomial term $\phi(B)\Phi(B^s)$ models the autoregressive part of the time series and $\theta(B)\Theta(B^s)$ stands for the moving average part. The presence of polynomial terms in B^s enables the seasonal dependence in the traffic flow to be modeled. Box and Jenkins' methodology [2] is used to estimate all the parameters characterizing a given time series. Model orders are fixed by analyzing the autocorrelation and partial autocorrelation functions of time series: the results obtained are summarized in section III.

From a practical point of view, the relevance of the SARIMA process to model traffic flows depends on the ability of the model to forecast values of the traffic flow. SARIMA time series can be represented in several forms: here, only the linear state-space form is retained. Indeed, as discussed in [1], this model has many virtues and especially the availability of Kalman recursions. Let x_t be the stationary AutoRegressive and Moving Average (ARMA) process verifying:

$$x_t = (1-B)^d(1-B^s)^D y_t \quad \text{and} \quad \phi(B)\Phi(B^s)x_t = \theta(B)\Theta(B^s)e_t.$$

Then, the derivation of the state-space model from the SARIMA process leads to (see details in [1, p. 471]):

$$\left\{ \begin{array}{rcl} y_t & = & F\mathbf{x}_t + e_t \\ \mathbf{x}_{t+1} & = & G\mathbf{x}_t + He_t \end{array} \right. ,$$

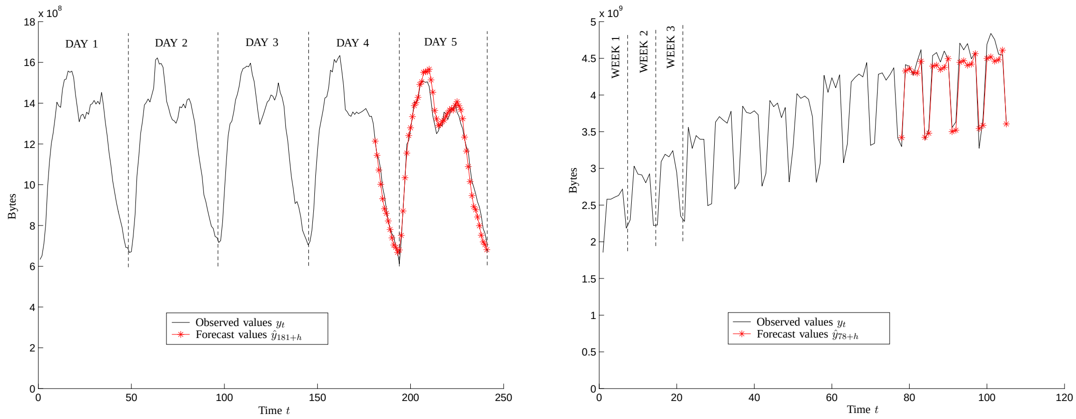
where F, G and H are matrices with appropriate dimensions depending on the parameters of the SARIMA model. The state vector $\mathbf{x}_t$, defined as $\mathbf{x}_t = (x_{t-m+1},\dots,x_t,y_{t-d-sD+1},\dots,y_t)^T$ with $m = \max\{p+sP, q+sQ\}$, is composed of 1) the past ARMA values $x_{t-m+1}, \ldots, x_t$ which stand for the stationary autoregressive $\phi(B)\Phi(B^s)$ and moving average $\theta(B)\Theta(B^s)$ part of the model and 2) past observations $y_{t-d-sD+1}, \ldots, y_t$ which stand for the non-stationary term $(1-B)^d(1-B^s)^D$.

Classical Kalman recursions are then used to forecast the traffic flow. In particular, at instant t, given the current and past observations $y_t, y_{t-1}, \ldots$, the Kalman h-step predictor gives the best linear prediction $\hat{y}_{t+h}$ of the value y_{t+h}. Kalman recursions permits an on-line calculation of $\hat{y}_{t+h}$.

III. Experimental results

To validate our model, a one-year traffic trace with a 5 minute observation interval is used. The trace was obtained from SNMP (Simple Network Management Protocol) reports of the routers freely accessible on the net and one of the most loaded routers representing the higher activity was selected. It can be seen from Fig. 1.(a,b) that the observed traffic has very noticeable daily and weekly periodicities and a tendency of traffic growth over the time. These periodicities correspond to human activity, and the observed trend is the proof that the traffic demand is increasing. The higher the bandwidth of the link, the higher the human activity and the more and more noticeable the periodicities become.

To illustrate the theoretical developments, we consider two levels of data aggregation. The time origin is assumed to be 0 for each case. On the one hand, SNMP measurements are aggregated over half an hour and only five days are analyzed (see Fig. 1.(a)). It can be shown that this traffic flow can be well approximated by a SARIMA $(1,1,1) \times (1,1,1)_{48}$ model and the noise e_t is approximately distributed as a zero mean white Gaussian noise with a well-estimated standard deviation $\sigma = 1.55 \times 10^8$. Since the fourth day in the morning, the Kalman predictor algorithm is used to forecast the traffic for a long lead time corresponding to the afternoon of the fourth day and the whole of the fifth day (dotted line with stars in Fig. 1.(a)). This simple model faithfully reproduces the seasonal pattern and supplies good forecasts. On the other hand, SNMP measurements are aggregated day by day for 15 weeks. Box and Jenkins' methodology [2] permits a SARIMA $(2,1,0) \times (1,1,0)_7$ model to be validated with $\sigma = 5.9 \times 10^7$. The ability of this model to forecast values of the traffic flow for the last four weeks is shown on Fig. 1.(b).



(a) (b)

Fig. 1. Traffic flow values measured every (a) $\Delta = 30$ minutes and (b) every day ($\Delta = 24$ hours).

IV. Conclusion

This paper shows that an Internet traffic flow can be well modeled by a SARIMA process by taking into account seasonal traffic patterns. In particular, Kalman recursions provides good forecasts. SARIMA models with multiple seasonal patterns will be studied in future work to consider traffic flows observed over a longer time period with a finer aggregation level.

References

[1] P. J. Brockwell and R. A. Davis, *Time series: theory and methods, second edition*. Springer, 1991.
[2] G. E. P. Box and G. M. Jenkins, *Time series analysis: forecasting and control, revised edition*. Holden-Day, 1976.