A Comparative Study of the Statistical Methods Suitable for Network Traffic Estimation

IARINA MARIAN, VASILE DADARLAT, BOGDAN IANCU
Computer Science Department
Technical University of Cluj-Napoca
26-28 G. Baritiu Street, Cluj-Napoca
ROMANIA

Abstract: Predicting network traffic has a great importance for many real time and non-real time applications, for planning the network resources and for traffic matrix computations. In this paper several estimation methods for IP network traffic are studied. Methods for both Short-Range Dependence (SRD) and Long-Range Dependence (LRD) characteristics are presented, and also offline and online prediction algorithms are described with their advantages and disadvantages and suitable practical applications.

Key-Words: Estimation Methods, Prediction, Network Traffic, Accuracy.

1 Introduction

Forecasting calculations are performed by the majority of telecommunications service providers, in order to assist them in planning their networks and services. This process helps ISPs, well in advance of product launch, to make key investment decisions relating to product development, introduction, and pricing.

Forecasting can be conducted for many reasons, so it is important that the purpose for performing the calculation is defined and understood. Some common reasons for using forecasting methods in networking are: to help network planners to plan and budget their resources, to provide some tool for the management to evaluate if the decisions that have been made will benefit the company and also to verify if the outcomes predicted by the old forecasts are confirmed by some new forecasts.

A variety of ways to conduct and to categorize forecasting methodologies were developed, for instance [1]: genius forecasting, trend extrapolation, consensus methods, simulation methods, cross-impact matrix method, scenario, decision trees.

Several methods have been used to improve and test forecast accuracy. It is difficult to determine the accuracy of any forecast, because it represents an attempt to predict future events. A simple method to check the accuracy involves the use of several different forecasting methods and comparing the results to see if they are relatively equal. Another method can involve statistically calculating the errors in the forecasting calculation and expressing them in terms of the root mean squared error, thus providing an indication of the overall error in the method. Determining forecast accuracy can never be performed with certainty, because it depends on the accuracy of the input data, on the selected forecasting method and on the forecasting process.

For network operation (e.g., traffic matrices computation, estimation of the quality of service perceived by end users, anomaly detection, etc.), it is crucial to understand to which extent sampled data reflect original traffic, in particular flow statistics (e.g., their number of packets or their duration, the number of flows, etc). One of the major problems, when performing packet sampling, consists of recovering original characteristics of traffic from sampled data.

Time series methods are based on measurements taken of events on a periodic basis. These methods use such data to develop models that can then be used to extrapolate into the future, thereby generating the forecast. Each model operates according to a different set of assumptions and is designed for a different purpose.

All of the studied methods are applied on estimating traffic over IP networks and their results in LAN and WAN networks. It is useful to remind also the gains from other methods applied in ATM networks as in [2], wireless networks as in [3], [4], [5] and also MPLS [6].
The second and third sections of the paper present the theoretical aspects and results of the studied methods. The fourth section presents the results comparison and makes a global analysis of the estimation methods. Finally, concluding remarks are given in the last section.

2 Theoretical Background

2.1 Short-Memory Models

A first class of short memory models is represented by regression models. Regression models are statistical models which describe how changes in one or more variables will change the value of another variable within a specified time window and a moving average of white noise.

In [7] and [8] a review of the main short memory models is presented: Autoregressive Model, Moving Average Model, Autoregressive Moving Average Model and Autoregressive Integrated Moving Average Model.

2.1.1 Autoregressive Model

The autoregressive model of order p, AR (p), has the form:

$$X_t = a_0 + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t, \ t>0$$  \hspace{1cm} (1)

where \(a_i\) are real constants and \(\varepsilon_t\) is white noise (independent identically distributed random variables with mean 0 and variance \(\sigma^2_\varepsilon\)).

AR models can be used to model stationary (time series that have a constant mean) and invertible time series.

2.1.2 Moving Average Model

The moving average model of order q, MA (q), has the form

$$X_t = \sum_{i=0}^{q} b_i \varepsilon_{t-i}, \ t>0$$  \hspace{1cm} (2)

where \(b_i\) are real constants and \(\varepsilon_t\) is the white noise.

2.1.3 Autoregressive Moving Average Model

The autoregressive moving average model of order (p,q) ARMA(p, q) has the form:

$$X_t = a_0 + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=0}^{q} b_i \varepsilon_{t-i}$$  \hspace{1cm} (3)

This model is more flexible than AR and MA in modeling time series, but it can only model non-stationary time series.

In practice, it is frequently true that adequate representation of actual time series can be obtained with models, in which \(p\) and \(q\) are not greater than 2 and often less than 2 [7].

2.1.4 Autoregressive Integrated Moving Average Model

The autoregressive integrated moving average model ARIMA (p, d, q) is an extension of ARMA (p, q) and can be used to model non-stationary processes. ARIMA (p, d, q) has the form:

$$\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t$$  \hspace{1cm} (4)

where \(B\) is the backshift operator (\(BX_t = X_{t-1}\), \(B^i X_t = X_{t-i}\)), \(\phi(B)\) is the autoregressive operator, \(\theta(B)\) is the moving average operator, \(d\) is the order of differencing (if \(d=0\), the model is equivalent to an ARMA model) and \(\varepsilon_t\) is the white noise.

ARMA and ARIMA are short range dependent models and are not considered suitable for modern high-speed networks.

A second class of short memory models is represented by Markov models. Markov analysis looks at a sequence of events (states), and analyzes the tendency of one event (state) to be followed by another. Using this analysis, a new sequence of random but related events can be generated, which will look similar to the original. A Markov process is useful for analyzing dependent random events - that is, events whose likelihood depends on what happened last. A stochastic process is said to have the Markov property if the conditional probability distribution of future states of the process, given the present state, depends only upon the current state. In general, increasing the number of states results in a more accurate model at the expense of increased computational complexity. Markov-type models often result in a complicated structure and many parameters, when used to model a long-range dependent process [7].

2.2 Long-Memory Models

2.2.1 Fractional ARIMA Model (FARIMA)
Fractional ARIMA is a natural generalization of the ARIMA (p, d, q) process, by allowing real values for differencing parameter d, instead of only integer values [9]. \( X_t \) is a stationary invertible FARIMA (p, d, q) process if:

\[
\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t,
\]

where \( B \) is the backshift operator, \( \phi(B) \) is the autoregressive operator, \( \theta(B) \) is the moving average operator, \( d \) is a real number, \( d \in (-0.5,0.5) \) - the order of differencing, and \( \varepsilon_t \) is the white noise.

### 2.2.2 Generalized ARMA Model (GARMA)

In order to model both short-range and long-range dependencies in a time series, a generalization of all the regression models was introduced: GARMA model [10]. GARMA(p,q) has the form:

\[
\phi(B)(1-2\eta B + B^2)^d X_t = \theta(B)\varepsilon_t
\]

where \( B \) is the backshift operator, \( \phi(B) \) is the autoregressive operator, \( \theta(B) \) is the moving average operator, \( d \) is a real number, \( d \in (-0.5,0.5) \) - the order of differencing, \( \eta \in (-1,1) \), \( (1-2\eta B + B^2)^d \) is the Gegenbauer polynomial and \( \varepsilon_t \) is the white noise.

### 2.2.3 Fractional Brownian Motion (FBM) and Fractional Gaussian Noise (FGN)

Fractional Brownian motion is also known as the Random Walk Process. It basically consists of steps in a random direction and with a step-length that has some characteristic value. The Fractional Brownian Motion (FBM) is a self-similar process with self-similarity parameter H, H being the Hurst parameter.

FBM is useful for theoretical analysis, in practice fractional Gaussian noise (FGN) is often used. FGN is a stationary process which exhibits long-range dependence with Hurst parameter H.

### 3 TRAFFIC ESTIMATORS

This section briefly discusses the main elements of several traffic estimators, pointing out, in a comparative manner, the advantages of the approaches, the problems solved and also their limitations.

A definition of Long-Range Dependence (LRD) found in [11], states that LRD is a statistical phenomenon observed in some time series. A time series which has LRD appears stationary overall, remains at higher or lower values than its mean for relatively long periods of time and appears to exhibit cycles or trends but with no clear overall cycle emerging. An important observation concerning Internet traffic, when observed across very long time scales, is that exhibits LRD.

On the other hand, short-range dependence (SRD) means that the behavior of a time-dependent process does not show statistically significant correlations across large time scales (current observations are not correlated to very old observations).

#### 3.1 Minimum Mean Square Error Predictor (MMSE)

For networks that can not be captured by traditional traffic models, one simple solution to the estimation problem is minimum mean square error (MMSE) described in [7], that requires matrix inversion and autocorrelation computation and in which a weight vector is derived by minimizing the expected value of squared errors.

MMSE predictor presents a series of advantages as the simplicity of implementation and the fact that it does not need to know the underlying structure of traffic (can be used for on-line traffic prediction).

In [7] the accuracy of several methods for forecasting network traffic is tested, using Ethernet traffic traces. The simulation outcomes, obtained by applying the traces on MMSE, and also on FARIMA, GARMA, FGN, showed the following results. First, increasing the history size beyond a limit will not increase the accuracy of the models, except for GFN. On the other hand this will increase the computation time.

Second, for MMSE, the optimal history size is much smaller than the ones for FARIMA and GARMA.

Third, FGN proved to be a pure LRD process, since the accuracy of the forecast increases with the history size.

There are several applications, both in theory and in practice, that need an exact model for the network traffic. In these situations models like FARIMA are the best choice, but if a simple online predictor, with an acceptable performance and accuracy is required, MMSE predictor is the most suitable predictor.

#### 3.2 ARIMA/GARCH

The Auto Regressive Integrated Moving Average (ARIMA) with Generalized Auto Regressive Conditional Heteroscedasticity (GARCH) model is a non-linear time
series model which combined the linear ARIMA with conditional variance GARCH. The model has the ability to capture both SRD and LRD characteristics. A comprehensive description of the ARIMA/GARCH model is found in [12].

In order to capture the bursty characteristic which has been found in data traffic, the ARIMA/GARCH model can be used.

Traditional time series prediction models, such as ARIMA cannot capture this characteristic, because the variance is constant. For this reason, GARCH model is introduced to explain the bursty characteristics. The most important contribution of GARCH model is its dynamic variance, where the variance varies over time.

In [12] the accuracy of the proposed model is tested comparatively to FARIMA, using TCP traffic traces.

The test results showed that the performance of the ARIMA/GARCH model outperforms the FARIMA model, by having superior adaptability and accuracy.

Due to its computation complexity ARIMA/GARCH model can be best used in offline prediction applications.

3.3 Le Cam’s Inequality and Chen Stein Method

A model for inferring flow statistics is proposed in [13] and consists of computing the probability distribution of the number of packets in sampled flows. The challenge is to extract information from the distribution in order to infer the statistics of original flows.

The approach considers an unnormalized sampled flow distribution – the number of flows which are sampled j or at least j times for j \( \geq 1 \). The random variables describe the number of flows seen j times or at least j times in successive time windows with length \( \Delta \) for \( j \geq 1 \). The characteristics of these random variables depend upon the statistics of flows. The authors use Le Cam’s inequality and Chen Stein method to prove that the size of a sampled flow can be obtained via a scaling of the original size of the flow.

The authors consider real traffic traces in order to test the theoretical results obtained via Le Cam’s and Chen Stein method.

Considering the number of flows - sampled at least j times, and Chen Stein method, it is shown in [13] that the probability distributions of the sampled flow size and the original flow size can be related when the sampling rate is small.

The main limitation of this approach is that the method used to estimate the number of sampled flows, will play a significant role for the estimation of original flow statistics. The problems arise from the fact that sampling alters information on small flows. By using a simple method based on numerous observation experiments of real Internet traffic, the authors have developed a method of inferring the number of flows by assuming that the number of packets in small flows is geometrically distributed. This allows recovering information on small flows and then to infer flow statistics with a reasonable accuracy.

The next section presents the results comparison and makes a global analysis of the previous discussed estimation methods.

4 Results Comparison

The studied methods show that non-linear time series model can model and forecast better than the classical linear time series models, even the linear time series model can also behave self-similarity. The basic idea, to introduce the non-linear time series model instead of the linear time series model, is based on the network’s property: traffic burstiness. In the statistical theory, if the time series behaves bursty, that means its variance fluctuate in time. In computer networks, burstiness is the characteristic that is determined by different outliers that influence network traffic making it more difficult to estimate.

The following table summarizes the characteristics of the studied estimators.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type of traffic</th>
<th>Memory model</th>
<th>Type of prediction</th>
<th>Dependence</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE</td>
<td>LAN</td>
<td>Short</td>
<td>Online</td>
<td>SRD</td>
<td>Low</td>
</tr>
<tr>
<td>ARIMA</td>
<td>LAN</td>
<td>Short</td>
<td>Offline</td>
<td>LRD</td>
<td>Medium</td>
</tr>
<tr>
<td>ARIMA/GARCH</td>
<td>WAN</td>
<td>Short/Long</td>
<td>Offline</td>
<td>LRD</td>
<td>High</td>
</tr>
<tr>
<td>Chen-Stein</td>
<td>WAN</td>
<td>Long</td>
<td>Offline</td>
<td>LRD</td>
<td>Medium</td>
</tr>
</tbody>
</table>

Table 1. Comparison of Estimators

MMSE is method that can be used as an on-line predictor while it is very simple to implement and very efficient for SRD traffic but it lacks the capability to estimate accurately traffic over a long period of time.

The ARIMA models capture the short-range dependence (SRD) characteristics only and thus are suitable for prediction traffic traces on short period of time. ARIMA is one of the methods used to estimate
traffic in wireless networks. The ARIMA/GARCH model is a non-linear time series model which can capture both SRD and LRD characteristics of traffic. When the number of predicted steps is small ARIMA/GARCH model is more accurate than FARIMA model but is more complex and unstable.

The Chen Stein method captures LRD characteristics of traffic and is also applied for larger network traffic traces.

The comparative approach gives the originality of the presented paper.

5 Future Work

Future work will involves the implementation of the several prediction methods described in this paper, which will be applied to the end-to-end QoS framework being developed on the ongoing QAF project and the comparison of the estimation errors of each method.

6 Conclusions

In this paper different methods to predict network traffic are studied. Both methods for SRD and LRD are presented, offline and online prediction algorithms with their advantages and disadvantages and suitable practical application.

MMSE based methods can be used as on-line predictors while they are very simple and efficient. Simulation results show that for the traffic trace studied in [7] the best history size is 100 samples for both FARIMA and GARMA predictors. An important observation is that optimal history size for MMSE is about 20 samples which is much smaller than the optimal history size of FARIMA and GARMA. The interesting observation is the difference between accuracy of complex off-line predictor FARIMA and simple on-line predictor MMSE, less than 5%.

The ARIMA/GARCH model proposed in [12] is a non-linear time series model which can capture the conditional variance (variance vary time) very well. ARIMA/GARCH combined model is better than FARIMA model when the number of predicted steps is small even if it is noted that ARIMA/GARCH prediction methodology is more complex and unstable.

By considering the number of flows and Chen Stein method, it is shown that the probability distributions of the sampled flow size and the original flow size can be related when the sampling rate is small. By using a simple method based on numerous observation experiments of real Internet traffic, the authors in [13] have developed a method of inferring the number of flows by assuming that the number of packets in small flows is geometrically distributed.

Concluding, the goal of the paper was to study the suitability of different statistical methods for the traffic estimation in various communication networks. By this comparison, it is possible to choose the appropriate method for each specific case.

Acknowledgments:

This work was supported by the ‘QAF - Quality of Service Aware Frameworks for Networks and Middleware’ research project within the framework ‘National Research, Development and Innovation Program’ initiated by The National University Research Council - Romania (CNCSIS).

References:


