Available online at www.sciencedirect.com
SCIENCE DIRECT
Performance Evaluation 58 (2004) 25–42
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

Filtering and forecasting problems for aggregate traffic in Internet links☆

T. Anjali [a], C. Bruni [b], D. Iacoviello [b], G. Koch [b], C. Scoglio [a,*]

[a] *Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA*
[b] *Department of Computer and Systems Science, University of Rome "La Sapienza", Via Eudossiana 18, 00184 Rome, Italy*

Received 30 July 2003; received in revised form 22 March 2004

Available online 19 July 2004

Abstract

An important problem in bandwidth allocation and reservation over a communication link is to estimate the traffic bit rate in that link. This can be done by using specific tools for measurements of the traffic bit rate. However, the obtained measures are affected by some noise. Moreover, one might be interested in future traffic forecasting, when a prediction is needed. In this paper, an iterative filtering procedure is proposed for updating the traffic estimate upon the arrival of a new measurement. A birth and death stochastic model is assumed for the traffic bit rate to provide dynamical equations for the average behavior in the absence of information carried by measurements. Approximate solutions of the same updating problem are also given under the assumption that the posterior distribution of the traffic bit rate belongs to a specific class (beta or Gaussian distribution). This leads to approximate filtering procedures, which are expected to provide significant computational advantages. Finally, results obtained by processing simulated and real data are presented; stressing that the practical behavior of the approximate filters is quite satisfactory.

Keywords: Internet traffic measurement; Measurement noise; Filtering; Forecasting

1. Introduction

Let us consider an Internet link (or in general, a telecommunication link) and the carried traffic on the link. To guarantee the end-to-end Quality of Service (QoS) requirements of an aggregate of requests, a given amount of bandwidth has to be reserved on the link. A bandwidth broker (BB) is the management tool that makes bilateral agreements with its neighboring BBs, to allocate and reserve the required bandwidth [13]. Since the bit rate of the request is variable, also the allocated bandwidth should be adapted to follow those variations. A proposed scheme for resource provisioning is to have a bandwidth cushion,

☆ The work of T. Anjali and C. Scoglio was supported by NSF under award number 0219829.
* Corresponding author. Tel.: +1-404-894-6616; fax: +1-404-894-7883.
E-mail address: caterina@ece.gatech.edu (C. Scoglio).

0166-5316/$ – see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.peva.2004.05.002

wherein extra bandwidth is reserved over the current usage, in order to overcome the problem of the exact evaluation of the traffic. As proposed in [15], if the traffic volume on a link exceeds a certain percentage of the agreement level, it leads to a multiplicative increase in the agreement. A similar strategy is proposed in case the traffic load falls below a considerable fraction of the reservation. This scheme satisfies the scalability requirement but leads to an inefficient resource usage.

A more efficient management policy could be based on the knowledge of the actual overall traffic bit rate. However, any direct measurements of the bit rate are affected by measurement noise. Such noise tends to obscure the underlying traffic state and can lead to erroneous conclusions in traffic flow analysis. Removing such noise from the collected data is therefore desirable and can be accomplished by applying filters. Moreover, one might be interested in a forecasting of future traffic, when a prediction is needed, for the future traffic behavior.

As an example, Kalman filter has been applied to flow control in high-speed networks. In [8], Kalman filter was used for state estimation in a packet-pair flow control mechanism. In [10], Kalman filter was used to predict traffic in a collection of VC sources in one VP of an ATM network.

The traffic estimation problem was explicitly considered in [2] where a Kalman filter was adopted with reference to an approximate model. The well known Kalman filter yields the optimal solution for a dynamical estimation (filtering) problem whenever both dynamic and output equations are linear and both state and observation noises are white, Gaussian, and purely additive. A second contribution in this line has been given in [1], where the filtering problem for the underlying traffic is formulated and solved with reference to a general birth and death stochastic model.

In this paper, we further investigate the model already proposed in [1] and develop two approximated solutions for the underlying traffic filtering problem by introducing suitable simplifying assumptions. The behavior of these approximate filters with respect to the exact filter as well as their robustness properties are tested against suitable simulated and real Internet traffic data. More precisely, in Sections 2 and 3 we recall the stochastic birth and death model which describes the evolution and measurement of the traffic on a telecommunication link, as well as the exact filtering procedure in its two phases of forecasting and updating. While in the forecasting phase the conditional mean value and variances propagate according to closed form equation, the updating phase involves full computation of the updated conditional distribution. If one wishes to waive this computational burden, it is necessary to approximate the conditional distribution itself with a member of a known distribution class. In Section 4 approximate solutions for the updating phase are given under the assumption that the posterior traffic distribution belongs to a specific class, namely beta or Gaussian. This leads to approximate filtering algorithms, which are expected to provide significant computational advantages. Note that the overall approximated algorithm which is derived under the above-mentioned Gaussian assumption differs from the one already presented in [2] in that the forecasting phase exactly follows the original birth-and-death model and the approximated Gaussian distribution used in the updating phase is built about the current state. Finally, Section 5 contains results from simulated and real data processing and some concluding remarks.

2. A model for the evolution and measurement of the traffic on a telecommunication link

Let $x(t) \in \{0, 1, \dots, N\}$ denote the number of active connections at time t in a given communication link, which allows a given maximum connection number N. A simple birth-and-death model for $x(t)$ may be given as follows [3,9,14]:

$$\mathrm{d}x(t) = \lambda(t)(N - x(t))\,\mathrm{d}t + [\mathrm{d}\nu_1(t) - \lambda(t)(N - x(t))\,\mathrm{d}t] - \mu(t)x(t)\,\mathrm{d}t - [\mathrm{d}\nu_2 - \mu(t)x(t)\,\mathrm{d}t], \quad (1)$$

where $\nu_1(t)$, $\nu_2(t)$ are doubly stochastic independent Poisson processes with rate processes $\lambda(t)(N-x(t))$ and $\mu(t)x(t)$, respectively, $\lambda(t)$ and $\mu(t)$ being the birth and death rates assumed to be known non-negative integrable functions. Characteristics of IP traffic at packet level are notoriously complex (self-similar). However, this complexity derives from much simpler flow level characteristics. When the user population is large, and each user contributes a small portion of the overall traffic, independence naturally leads to a Poisson arrival process for flows [8,9]. Real traffic traces were obtained from the ABILENE and our filters are tested for these traffic traces, showing the validity of the assumptions.

Eq. (1) may be solved starting from any given initial condition $x(\bar{t})$, $\bar{t} \le t$.

Let us denote by $p_k(t|\bar{t}) = P(x(t) = k|\bar{t})$ the probability that $x(t)$ equals k, conditioned upon all possible given information up to time $\bar{t}$. Such conditioning will be explicitly specified in the notation whenever needed. The birth-and-death model (Eq. (1)) corresponds to the following master equations for p_k:

$$\dot{p}_k(t|\bar{t}) = \lambda(t)[N-(k-1)]p_{k-1}(t|\bar{t}) + \mu(t)(k+1)p_{k+1}(t|\bar{t}) - [\lambda(t)(N-k) + \mu(t)k]p_k(t|\bar{t}), \quad k = 0, 1, \ldots, N, \quad (2)$$

where we set

$$p_{-1}(t|\bar{t}) = p_{N+1}(t|\bar{t}) = 0.$$

In a vector notation, Eq. (2) becomes

$$\dot{p}(t|\bar{t}) = Q(t)p(t|\bar{t}), \quad (3)$$

where

$$p(t|\bar{t}) = (p_0(t|\bar{t}) \cdots p_N(t|\bar{t}))^{\mathrm{T}}$$

$$Q(t) = \begin{pmatrix} -N\lambda(t) & \mu(t) & 0 & \cdots & \cdots & 0 \\ N\lambda(t) & -[(N-1)\lambda(t)+\mu(t)] & 2\mu(t) & 0 & \cdots & 0 \\ 0 & (N-1)\lambda(t) & -[(N-2)\lambda(t)+2\mu(t)] & 3\mu(t) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 2\lambda(t) & -[\lambda(t)+(N-1)\mu(t)] & N\mu(t) \\ 0 & \cdots & \cdots & 0 & \lambda(t) & -N\mu(t) \end{pmatrix}. \quad (4)$$

As well known, the solution of Eq. (3), for a fixed value $p(\bar{t}|\bar{t})$, is

$$p(t|\bar{t}) = \Phi(t,\bar{t})p(\bar{t}|\bar{t}), \quad (5)$$

where $\Phi(t,\bar{t})$ is the transition matrix which solves the equation

$$\frac{\partial \Phi(t,\bar{t})}{\partial t} = Q(t)\Phi(t,\bar{t}), \qquad \Phi(\bar{t},\bar{t}) = I. \quad (6)$$

From Eq. (3), a dynamical model for the mean value $E(x(t)|\bar{t})$ of $x(t)$, conditioned upon the available information up to $\bar{t}$, can be derived by simple pre-multiplication by the row vector L^{T}:

$$L^{\mathrm{T}} = (\,0 \quad 1 \quad 2 \quad \ldots \quad N\,),$$

$$\frac{\mathrm{d}E(x(t)|\bar{t})}{\mathrm{d}t} = L^{\mathrm{T}}\dot{p}(t|\bar{t}) = L^{\mathrm{T}}Q(t)p(t|\bar{t}) = -(\lambda(t)+\mu(t))E(x(t)|\bar{t}) + \lambda(t)N, \tag{7}$$

hence

$$E(x(t)|\bar{t}) = E(x(\bar{t})|\bar{t})\,\mathrm{e}^{-\int_{\bar{t}}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s} + \int_{\bar{t}}^{t}\mathrm{e}^{-\int_{u}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s}\lambda(u)N\,\mathrm{d}u. \tag{8}$$

Similarly, by pre-multiplying Eq. (3) by the row vector M^{T}:

$$M^{\mathrm{T}} = (\,0 \quad 1 \quad 4 \quad \ldots \quad N^2\,),$$

we get for the mean value of $x^2(t)$.

$$\begin{aligned}\frac{\mathrm{d}E(x^2(t)|\bar{t})}{\mathrm{d}t} &= M^{\mathrm{T}}\dot{p}(t|\bar{t}) = M^{\mathrm{T}}Q(t)p(t|\bar{t})\\ &= -2(\lambda(t)+\mu(t))E(x^2(t)|\bar{t}) + (\lambda(t)(2N-1)+\mu(t))E(x(t)|\bar{t}) + \lambda(t)N.\end{aligned} \tag{9}$$

Introducing the variance of $x(t)$:

$$\sigma^2(t|\bar{t}) = E((x(t)-E(x(t)|\bar{t})^2|\bar{t}) = E(x^2(t)|\bar{t}) - E^2(x(t)|\bar{t}).$$

Eq. (9) leads to the dynamical representation

$$\frac{\mathrm{d}\sigma^2(t|\bar{t})}{\mathrm{d}t} = -2(\lambda(t)+\mu(t))\sigma^2(t|\bar{t}) - (\lambda(t)-\mu(t))E(x(t)|\bar{t}) + \lambda(t)N \tag{10}$$

whose solution clearly is

$$\sigma^2(t|\bar{t}) = \sigma^2(\bar{t}|\bar{t})\,\mathrm{e}^{-2\int_{\bar{t}}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s} + \int_{\bar{t}}^{t}\mathrm{e}^{-2\int_{u}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s}[\lambda(u)N - (\lambda(u)-\mu(u))E(x(u)|\bar{t})]\,\mathrm{d}u. \tag{11}$$

Eqs. (5), (8) and (11) may be (and will be) used to get predictive probability distribution (as well as mean value and variance) of $x(t)$, $t \geq \bar{t}$ conditioned upon any given information available up to time $\bar{t}$.

It is easily seen that Eqs. (7)–(10) can be also directly obtained from 1 and the corresponding model for $x^2(t)$ by taking mean values of both sides.

All connections are supposed to employ the same known bandwidth C. A bandwidth broker is naturally interested in knowing the total bandwidth request $Cx(t)$. To that purpose, at the discrete times iT, $i = 0, 1, \ldots$ where $T > 0$ is a fixed sampling time, a specific device yields a measurement $y(iT)$ of $Cx(iT)$ which is affected by error $n(iT)$:

$$y(iT) = Cx(iT) + n(iT), \quad i = 0, 1, \ldots \tag{12}$$

The selection of the sampling time interval has to be performed in order to balance two opposing requirements. On one side, T has to be short enough to closely follow the traffic variations. On the other side, T

has to be long enough to limit the control effort. The choice of T depends on the relative importance of the above two requirements and on the traffic variability.

We concentrate, without loss of generality, on a single streaming traffic class. Streaming traffic classes are characterized by an inherent bit rate C that must be preserved in the network. To consider a scenario with different traffic classes having different bandwidth requirements (different values of C), the same analysis can be extended and applied for each class. The constant resource requirement can be considered valid in our context of DiffServ Expedited Forwarding class. Voice calls can be considered in this class, but also video connections belong to this class and they have very different values of C.

As far as the measurement noise is concerned, for security and privacy reasons, information in the routers cannot always be accessed. This is why a large effort in the scientific community has been devolved to the design of methods for the measurement of available bandwidth [5–7]. Available bandwidth is complementary to the utilization, so measuring available bandwidth is equivalent to measuring the aggregate traffic on the link. However, these methods are active measurements and thus intrinsically not exact. They are based on the transmission of packets at increasing speeds, up to a maximum which is taken as the value of the available bandwidth. This is the main basis for the belief that traffic measurements are noisy and a filter has to be used to estimate the real value of the aggregate traffic. However, even in the case in which router information can be accessed, the use of a tool to retrieve the information from the routers is still required. One of the most popular tools is Multi Router Traffic Grapher (MRTG) [12], which we have used for obtaining the real data in our performance evaluation section. Compared with the previous methods based on active measurement, MRTG provides more accurate measurements but still delays and errors in the transfer of the information can occur.

The sequence $\{n(iT), i = 0, 1, 2, \dots\}$ is assumed to be white. Each error sample $n(iT)$ is such that $y(iT) \in \{0, 1, \dots, y_M\}$ where $y_M = CN$. Besides, the same is probabilistically characterized by the values $q_h(iT|k)$ defined as follows:

$$q_h(iT|k) = P(y(iT) = h|x(iT) = k), \quad h \in \{0, 1, \dots, y_M\}, k \in \{0, 1, \dots, N\}. \tag{13}$$

At this point, the problem arises of optimally processing the available measurement data in order to get an estimate of $x(t)$.

3. Optimal filtering for the number of active connections

We are now ready to formulate the problem of filtering for $x(t)$, $t \in [iT, (i+1)T)$, that is the on-line iterative determination of the optimal estimate $\hat{x}(t|i)$ of $x(t)$, given all available information $y(0), y(T), \dots, y(iT)$ up to time iT (denoted for simplicity by $y_i = \{y(0), y(T), \dots y(iT)\}$).

Once specified that the available information is represented by the measurement values, we explicitly denote by $p_k(t|i)$, $iT \le t \le (i+1)T$ the probability that $x(t)$ takes on the value k given y_i:

$$p_k(t|i) = P(x(t) = k|y_i) \tag{14}$$

and by $p(t|i)$ the vector

$$p(t|i) = (p_0(t|i)p_1(t|i) \cdots p_N(t|i))^{\mathrm{T}}. \tag{15}$$

The above probability can be iteratively computed in two steps:

(1) the predictive step, that is the computation of $p(t|i)$ from $p(iT|i)$, $iT \le t \le (i+1)T$;

(2) the updating step (or innovation), that is the computation of $p((i+1)T|i+1)$ from $p((i+1)T|i)$.

The first step is already solved by Eq. (5) and uses the dynamical model of Eq. (3) for the free evolution of the distribution of $x(t)$ over the time interval $[iT, (i+1)T]$ with no new information besides the information already available at iT, that is y_i

$$p(t|i) = \Phi(t, iT)p(iT|i), \quad iT \leq t \leq (i+1)T$$

and in particular

$$p((i+1)T|i) = \Phi((i+1)T, iT)p(iT|i). \tag{16}$$

As far as the second step is concerned, by Bayes' formula and taking whiteness of $\{n(iT)\}$ into account, for $y((i+1)T) = h$, we get

$$\begin{aligned} p_k((i+1)T|i+1) &= \frac{P(y((i+1)T) = h|x((i+1)T) = k, y_i)p_k((i+1)T|i)}{P(y((i+1)T) = h|y_i)} \\ &= \frac{P(y((i+1)T) = h|x((i+1)T) = k)p_k((i+1)T|i)}{\sum_{l=0}^{N} P(y((i+1)T) = h|x((i+1)T) = l)p_l((i+1)T|i)} \\ &= \frac{q_h(i+1|k)p_k((i+1)T|i)}{\sum_{l=0}^{N} q_h(i+1|l)p_l((i+1)T|i)}. \end{aligned} \tag{17}$$

By introducing the matrix

$$U_h(i+1) = \operatorname*{diag}_{0\leq k\leq N}\{q_h(i+1|k)\}. \tag{18}$$

Eq. (17) can be written in vector notation as

$$p((i+1)T|i+1) = \frac{U_h(i+1)p((i+1)T|i)}{1^{\mathrm{T}}U_h(i+1)p((i+1)T|i)}, \tag{19}$$

where $1^{\mathrm{T}} = (1 \quad 1 \quad \cdots \quad 1)$.

Eqs. (16) and (19) allow us to compute $p((i+1)T|i+1)$ from $p(iT|i)$.

The optimal estimate $\hat{x}(t|i)$ of $x(t), iT \leq t < (i+1)T$ is achieved by considering the minimum conditional variance criterion of the estimate error, and thus it is given as

$$\hat{x}(t|i) = E(x(t)|y_i) = L^{\mathrm{T}}p(t|i).$$

Since the estimation error

$$e(t|i) = x(t) - \hat{x}(t|i)$$

has zero mean value (i.e. the estimate $\hat{x}(t|i)$ is unbiased), the variance $\sigma^2(t|i)$ of $e(t|i)$ is simply given by

$$\sigma^2(t|i) = M^{\mathrm{T}}p(t|i) - \hat{x}^2(t|i).$$

Obviously, even the computation of the optimal estimate and its variance can be performed in two steps: prediction and update.

In the prediction step, similarly to what happens for the conditional probability, the iterative structure still holds for both conditional mean value and conditional variance of the estimation error. In fact, for

$t \in [iT, (i+1)T)$, they evolve, according to Eqs. (8) and (11), as solutions of the linear differential equations (7) and (10), respectively, keeping the following expressions:

$$\hat{x}(t|i) = \hat{x}(iT|i)\,\mathrm{e}^{-\int_{iT}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s} + \int_{iT}^{t} \mathrm{e}^{-\int_{u}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s}\lambda(u)N\,\mathrm{d}u, \tag{20}$$

$$\sigma^2(t|i) = \sigma^2(iT|i)\,\mathrm{e}^{-2\int_{iT}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s} + \int_{iT}^{t} \mathrm{e}^{-2\int_{u}^{t}(\lambda(s)+\mu(s))\,\mathrm{d}s}[\lambda(u)N - (\lambda(u)-\mu(u))\hat{x}(u|i)]\,\mathrm{d}u. \tag{21}$$

On the contrary, in the update step the iterative structure vanishes, as far as the conditional mean and its variance are concerned. In fact, at the measurement times, the conditional distribution undergoes a discontinuity as described by (19); as a consequence, conditional mean value and variance also exhibit a discontinuous behavior. In particular, we have

$$\hat{x}((i+1)T|i+1) = \frac{L^{\mathrm{T}}U_h(i+1)p((i+1)T|i)}{1^{\mathrm{T}}U_h(i+1)p((i+1)T|i)}, \tag{22}$$

$$\sigma^2((i+1)T|i+1) = \frac{M^{\mathrm{T}}U_h(i+1)p((i+1)T|i)}{1^{\mathrm{T}}U_h(i+1)p((i+1)T|i)} - \hat{x}^2((i+1)T|i+1) \tag{23}$$

which require the knowledge of $p((i+1)T|i)$.

4. Approximate filters

In the previous section, we have seen that computation of predictive conditional mean value $\hat{x}(t|i)$ and conditional variance $\sigma^2(t|i)$ of the estimation error can be performed according to Eqs. (20) and (21), respectively.

The only really consistent computational burden is related to the update of the estimation through the innovation step occurring at measurement times $(i+1)T$. Indeed, as already mentioned, computation of $\hat{x}((i+1)T|i+1)$ and $\sigma^2((i+1)T|i+1)$ (according to Eqs. (22) and (23)) requires full knowledge of $p((i+1)T|i)$; this in turn requires solution of Eq. (6). The latter is a linear time varying differential equation system of dimension $N+1$.

The issue then arises of specific approximations, which allow direct update of $\hat{x}(t|i)$ and $\sigma^2(t|i)$ at time $t=(i+1)T$ without computation of $p((i+1)T|i)$.

For instance, one could think of suitable numerical approximations for $p_k((i+1)T|i)$ with respect to k, with the aim of transforming Eqs. (22) and (23) into a system of iterative equations in the pair $\hat{x}(iT|i)$ and $\sigma^2(iT|i)$.

A technically sounder approach is to introduce structural modifications in the model suitable to guarantee that $p((i+1)T|i)$ be uniquely defined by its mean value and variance, and that this property is kept through the innovation step. This means assuming for the distribution of $x(iT)$ conditional on y_i and for the distribution of $y(iT)$, conditioned on $x(iT)$, suitable conjugate forms [4]. One possibility in this line is to assume for $x(t)/N$, conditioned upon y_i, $t \in [iT, (i+1)T)$ a beta distribution, $\mathrm{Beta}(\alpha(t|i), \beta(t|i))$ and for $y(iT)$, conditioned upon $x(iT)$, a binomial distribution $\mathrm{Bin}(y_M, Cx(i)/y_M)$.

In the following the solution obtained by these assumptions will be referred to as "binomial filter". As known [4]:

$$p_{x(t)/N}(\xi|y_i) = K(\alpha(t|i), \beta(t|i))\xi^{\alpha(t|i)-1}(1-\xi)^{\beta(t|i)-1}, \quad \xi \in [0, 1], \tag{24}$$

where $K(\alpha(t|i), \beta(t|i))$ is a normalization factor (inverse of the beta function).

We must observe that the probabilistic nature assumed in Eq. (24) for $x(t)/N$ corresponds to a continuous behavior for $x(t)$ (no more constrained to assume integer values) but still confined in the range $[0, N]$.

Mean and variance of $x(t)$, $t \in [iT, (i+1)T)$, given y_i, are respectively given by

$$\hat{x}(t|i) = \frac{N\alpha(t|i)}{\alpha(t|i)+\beta(t|i)}, \tag{25}$$

$$\sigma^2(t|i) = \frac{N^2\alpha(t|i)\beta(t|i)}{(\alpha(t|i)+\beta(t|i))^2(\alpha(t|i)+\beta(t|i)+1)}. \tag{26}$$

As far as $y(iT)$ is concerned, we have

$$P(y(iT)=h|x(iT)) = \binom{y_M}{h}\left(\frac{x(iT)}{N}\right)^h\left(1-\frac{x(iT)}{N}\right)^{y_M-h}, \quad h \in \{0, 1, \ldots, y_M\}$$

with mean value and variance given by

$$E(y(iT)|x(iT)) = Cx(iT), \qquad \sigma^2(y(iT)|x(iT)) = Cx(iT)\left(1-\frac{x(iT)}{N}\right).$$

Eqs. (25) and (26) may be easily solved in terms of $\alpha(t|i)$ and $\beta(t|i)$:

$$\alpha(t|i) = \frac{\hat{x}^2(t|i)}{\sigma^2(t|i)} - \frac{\hat{x}^3(t|i)}{N\sigma^2(t|i)} - \frac{\hat{x}(t|i)}{N}, \qquad \beta(t|i) = \left(\frac{\hat{x}(t|i)}{\sigma^2(t|i)} - \frac{\hat{x}^2(t|i)}{N\sigma^2(t|i)} - \frac{1}{N}\right)(N-\hat{x}(t|i)).$$

At the measurement time $(i+1)T$ updating of α and β is easily achieved by exploiting the conjugate character of beta and binomial distributions:

$$\alpha((i+1)T|i+1) = \alpha((i+1)T|i) + y((i+1)T),$$
$$\beta((i+1)T|i+1) = \beta((i+1)T|i) + y_M - y((i+1)T).$$

This allows to perform the innovation step without the use of full knowledge of $p((i+1)T|i)$:

$$\hat{x}((i+1)T|i+1) = \hat{x}((i+1)T|i) + \frac{N\sigma^2((i+1)T|i)[y((i+1)T) - C\hat{x}((i+1)T|i)]}{(CN-1)\sigma^2((i+1)T|i) + N\hat{x}((i+1)T|i) - \hat{x}^2((i+1)T|i)}, \tag{27}$$

$$\begin{aligned}
&\sigma^2((i+1)T|i+1)\\
&\quad = [\sigma^2((i+1)T|i)[N\hat{x}^2((i+1)T|i) - \hat{x}^3((i+1)T|i) - \hat{x}((i+1)T|i)\sigma^2((i+1)T|i)\\
&\qquad + Ny((i+1)T)\sigma^2((i+1)T|i)]][N\hat{x}((i+1)T|i) - \hat{x}^2((i+1)T|i) + (CN-1)\sigma^2((i+1)T|i)]^{-2}\\
&\qquad \times [N\hat{x}((i+1)T|i) - \hat{x}^2((i+1)T|i) - \sigma^2((i+1)T|i)](N-\hat{x}((i+1)T|i))\\
&\qquad + N\sigma^2((i+1)T|i)(CN - y((i+1)T))][N\hat{x}((i+1)T|i) - \hat{x}^2((i+1)T|i) + CN\sigma^2((i+1)T|i)]^{-1}.
\end{aligned} \tag{28}$$

Eqs. (27) and (28), along with the free evolutions of Eqs. (20) and (21), provide a constructive two-dimensional solution of the filtering problem.

A second possible choice for conjugate distributions of, conditioned on y_i and $x(iT)$ respectively, is to assume Gaussian distribution for both of them. This not only cancels the discrete character of $x(iT)$ and $y(iT)$, but also broadens their support (naturally positive and bounded) to the whole of R^1. Furthermore, differently from previous model, the distribution of the measurement error $n(iT)$ is now independent of $x(iT)$. The filter obtained with these assumptions will be called Gaussian filter.

On the other hand, the Gaussian assumption leads to the well-known Kalman–Bucy filter. Now, the innovation step is described by the classical equations:

$$\hat{x}((i+1)T|i+1) = \hat{x}((i+1)T|i) + K(i+1)[y((i+1)T) - C\hat{x}((i+1)T|i)],$$
$$\sigma^2((i+1)T|i+1) = (1 - CK(i+1))\sigma^2((i+1)T|i),$$

where the innovation gain $K(i+1)$ is given by

$$K(i+1) = \frac{C\sigma^2((i+1)T|i)}{C^2\sigma^2((i+1)T|i) + \sigma_n^2((i+1)T)} \tag{29}$$

In (29), $\sigma_n^2((i+1)T)$ denotes the value of the variance of the measurement error $n((i+1)T)$.

As far as the prediction step is concerned, Eqs. (20) and (21) still hold, so that again we achieve a constructive two-dimensional solution.

A final remark concerns the free evolution stochastic model leading to the prediction step of Eqs. (20) and (21). Instead of the model in Eq. (1) which corresponds to a discrete valued state, one might adopt a continuous valued state model such as

$$\mathrm{d}x(t) = \lambda(t)(N - x(t))\mathrm{d}t - \mu(t)x(t)\mathrm{d}t + \mathrm{d}\omega(t),$$

where the martingale $\omega(t)$ has zero mean and (state dependent) incremental variance $\sigma_\omega^2(t)\,\mathrm{d}t$ equal to the incremental variance of the martingale term in Eq. (1), that is

$$\sigma_\omega^2(t) = \lambda(t)(N - x(t)) + \mu(t)x(t).$$

This is the model approach taken in [2].

5. Performance evaluation and comparison of the filters

In order to assess the performance of the various filters proposed in the previous sections, as well as to compare their behaviors, we tested the filters both on simulated and real data. Results on simulated data are reported in Figs. 1–5. In all cases, the initial condition was assumed to be $x(0) = 25$ and data are shown for the long-range behavior $\{iT\}$, $i = 100, \ldots, 1000$. The initial transient time is not reported, because of its limited applicative interest. However, we noted that the behavior of all filters was quite satisfactory even in that time phase.

In the first three experiments, the rates λ and μ were taken time invariant, and the distribution of the measurement noise was chosen according to different models. In particular, in Fig. 3 the variance of the measurement noise corresponds to the variance assumed in the binomial filter when computed around the long range value of $x(t)$. In the last two experiments the rates λ and μ were assumed to vary in time

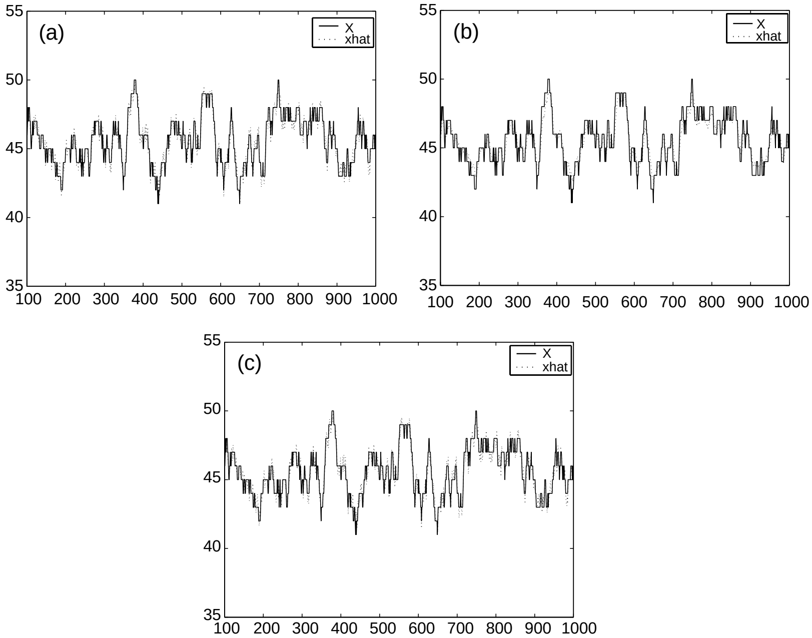

Fig. 1. Simulated data. $N = 50$; $C = 1$; $\lambda = 0.05$; $\mu = 0.005$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

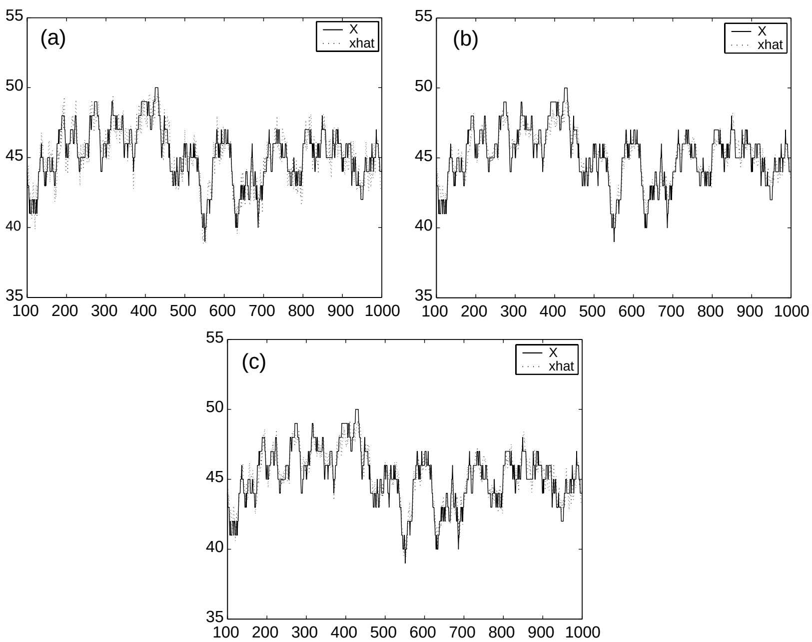

Fig. 2. Simulated data. $N = 50$; $C = 1$; $\lambda = 0.05$; $\mu = 0.005$; $n(iT) = -2$ (with prob. 0.2), -1 (with prob. 0.2), 0 (with prob. 0.2), 1 (with prob. 0.2), 2 (with prob. 0.2).

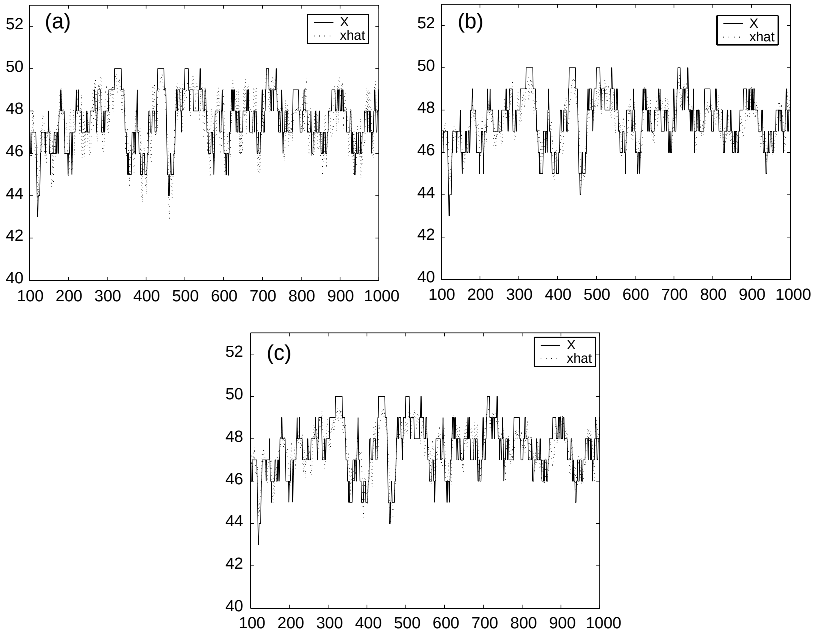

Fig. 3. Simulated data. $N = 50$; $C = 1$; $\lambda = 0.067$; $\mu = 0.003$; $n(iT) = -2$ (with prob. 0.2), -1 (with prob. 0.2), 0 (with prob. 0.2), 1 (with prob. 0.2), 2 (with prob. 0.2).

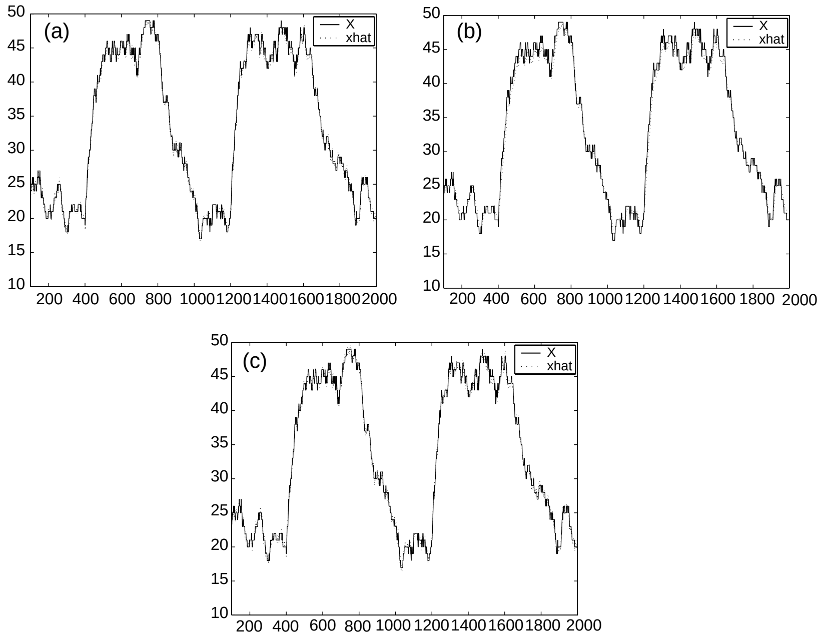

Fig. 4. Simulated data. $N = 50$; $C = 1$; $\lambda(t) = 0.005, t \in [0, 400] \cup (800, 1200] \cup (1600, 2000], 0.05, t \in (400, 800] \cup (1200, 1600]$; $\mu = 0.005$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

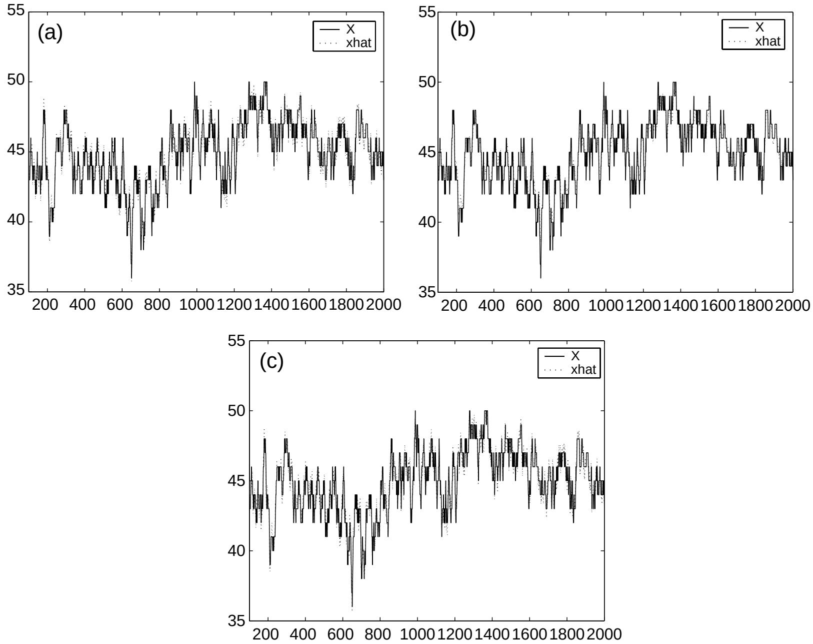

Fig. 5. Simulated data. $N = 50$; $C = 1$; $\lambda(t) = 0.05, t \in [0, 800] \cup (1600, 2000], 0.08, t \in (800, 1600]$; $\mu(t) = 0.005, t \in [0, 400] \cup (1200, 2000], 0.01, t \in (400, 1200]$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

according to the behavior reported in the captions. As a performance index, the coefficient of variation (CV) was computed in all experiments according to the formula:

$$\mathrm{CV} = \frac{\sqrt{(1/N)\sum_{i=100}^{1000}[x(iT) - \hat{x}(iT|i)]^2}}{(1/N)\sum_{i=100}^{1000} x(iT)}.$$

Based on the results of simulated data processing, the following remarks appear to be appropriate:

(1) The exact filter is able to closely follow the behavior of $x(t)$ in a very satisfactory way.
(2) Both approximate filters exhibit a performance not significantly worse than the exact one.

These remarks are robust against variations in the λ, μ, N values and/or variation in the measurement noise distribution.

We also applied the proposed filters to real data, with known C value (Figs. 6–9). Now a preliminary estimation of λ, μ, N was requested. We have estimated N according to the standard estimation procedure for the maximum admissible value of a random variable [11]. The values of λ and μ were derived observing that the average inter-arrival time is $(\lambda + \mu)/N\lambda\mu$ and the average inter-departure time is $(\lambda + \mu)/N\mu^2$. Parameters λ and μ have been estimated by evaluating the above times on the available historical data. Obviously, these estimation procedures work in the assumption of constancy of the unknown parameters. Should they vary in time, a joint state and parameter estimation problem arises. This is a complex issue indeed which could be the object of a future research.

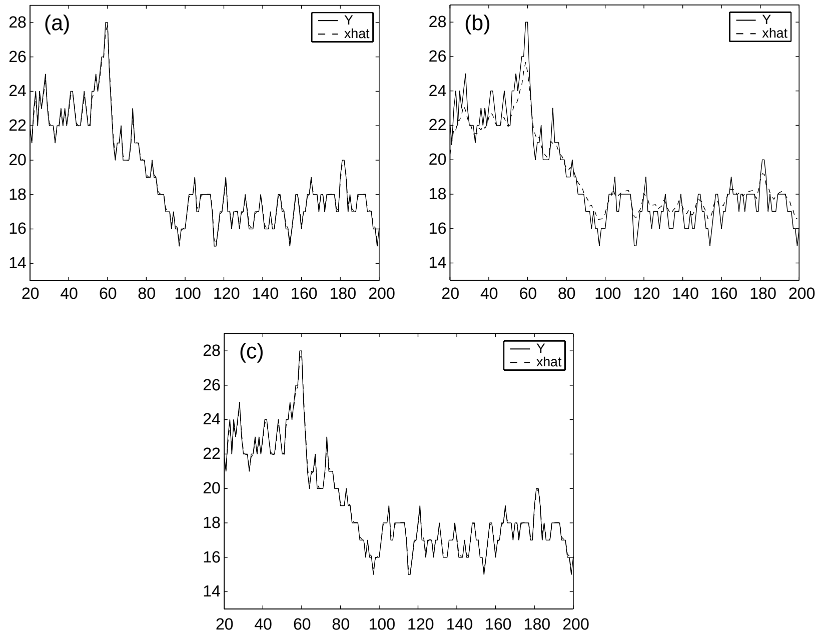

Fig. 6. Input traffic on nordunet interface of Abilene router in NYC on 20 May 2003 with $N = 35$; $C = 1$; $\lambda = 0.05$; $\mu = 0.04$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

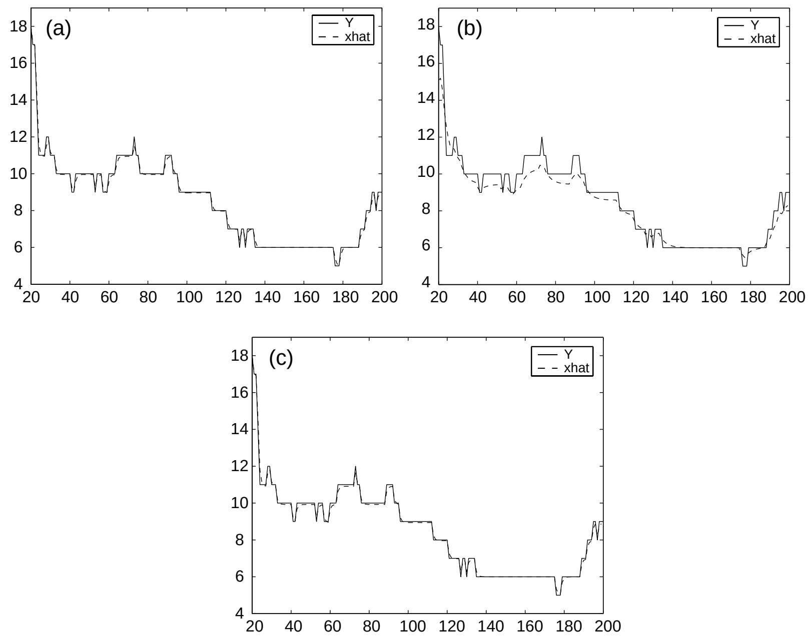

Fig. 7. Output traffic on nordunet interface of Abilene router in NYC on 20 May 2003 with $N = 30$; $C = 1$; $\lambda = 0.01$; $\mu = 0.04$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

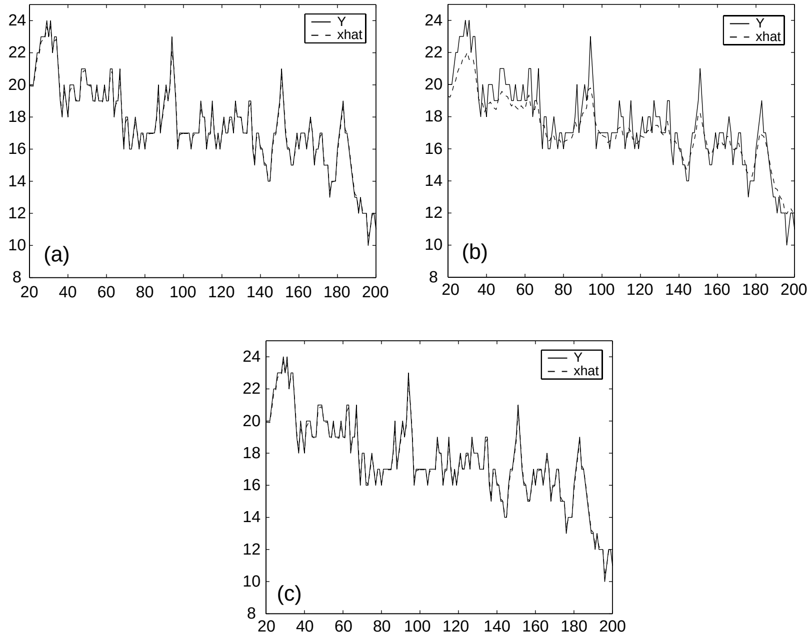

Fig. 8. Input traffic on SOX interface of Abilene router in Atlanta on 20 May 2003 with $N = 30; C = 1; \lambda = 0.05; \mu = 0.05; n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

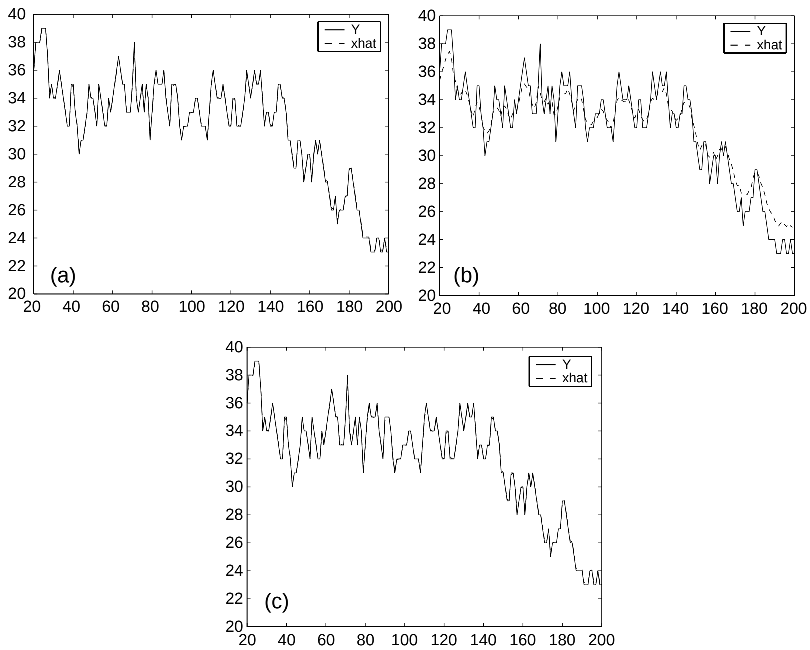

Fig. 9. Output traffic on SOX interface of Abilene router in Atlanta on 20 May 2003 with $N = 65; C = 1; \lambda = 0.05; \mu = 0.05; n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

We tested the above estimation procedures on simulated data with satisfactory results. As a matter of fact, also the distribution of the measurement noise n should be identified; however, being the estimation procedures robust against modifications of this distribution, we have assumed an a priori reasonable choice.

The performance of the approximated (binomial or Gaussian) filters as compared to the exact filter was assessed according to the following index:

$$\overline{\mathrm{CV}} = \frac{\sqrt{(1/N)\sum_{i=20}^{200}[\hat{x}_{\mathrm{exact}}(iT|i) - \hat{x}_{\mathrm{approx}}(iT|i)]^2}}{(1/N)\sum_{i=20}^{200}\hat{x}_{\mathrm{exact}}(iT|i)}.$$

Lastly, we wished to test robustness of filters against possible uncertainty in the values of parameters λ, μ, N (Figs. 10–12). The quantitative assessment of the robustness uses the $\underline{\mathrm{CV}}$ index defined as

$$\underline{\mathrm{CV}} = \frac{\sqrt{(1/N)\sum_{i=100}^{1000}[\hat{x}(iT|i) - \hat{x}'(iT|i)]^2}}{(1/N)\sum_{i=100}^{1000}\hat{x}(iT|i)},$$

where $\hat{x}(iT|i)$ denotes the filter output in the case of full knowledge of parameters values, while $\hat{x}'(iT|i)$ denotes the same output when an error is introduced in the parameter values themselves. Some conclusions may be drawn from the above:

(1) The optimal filter allows to iteratively compute the conditional means and variances by a closed form algorithm in the prediction step (see Eqs. (20) and (21)).

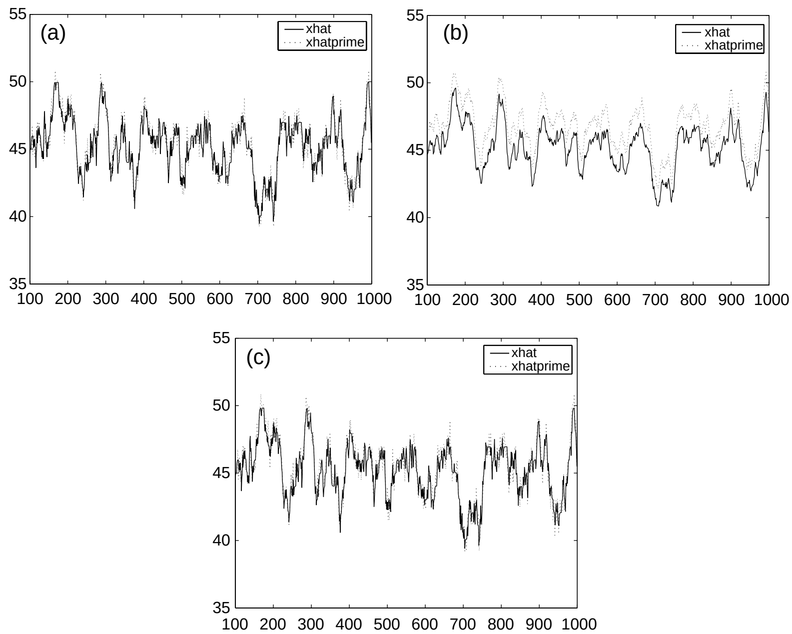

Fig. 10. Robustness analysis by simulated data; exact parameter values: $N = 50, C = 1, \lambda = 0.05, \mu = 0.005$; modified parameter values: $N = 60, C = 1, \lambda = 0.06, \mu = 0.006$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

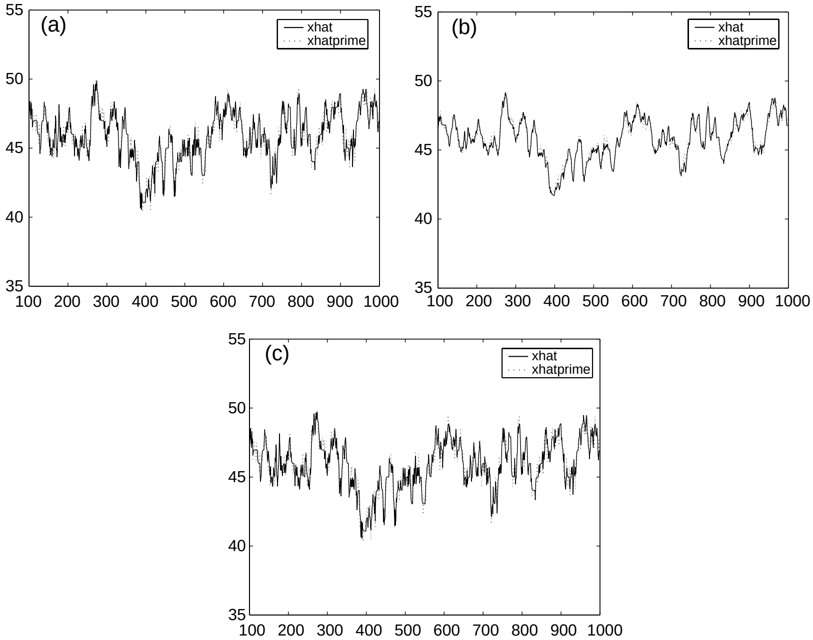

Fig. 11. Robustness analysis by simulated data; exact parameter values: $N = 50, C = 1, \lambda = 0.05, \mu = 0.005$; modified parameter values: $\lambda = 0.075$, $\mu = 0.0075$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

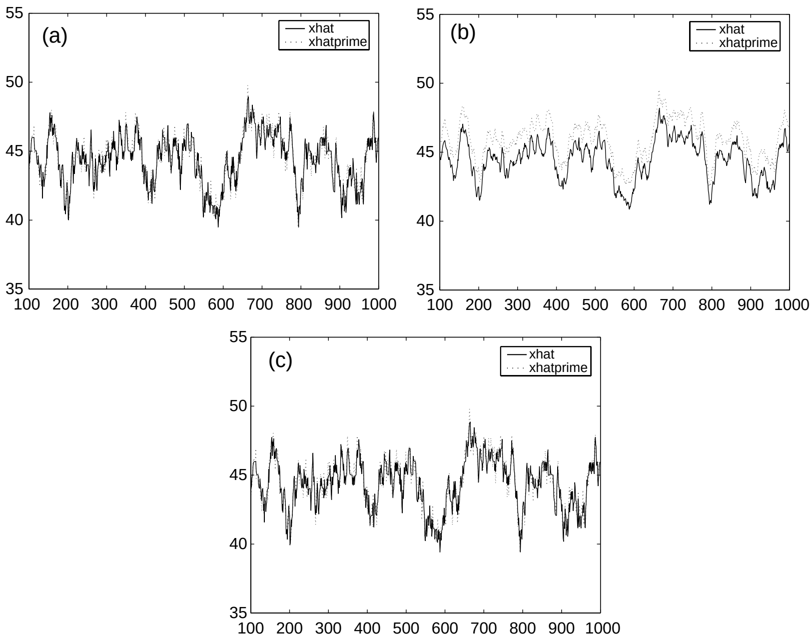

Fig. 12. Robustness analysis by simulated data; exact parameter values: $N = 50, C = 1, \lambda = 0.05, \mu = 0.005$; modified parameter values: $N = 60$; $n(iT) = -1$ (with prob. 0.15), 0 (with prob. 0.7), 1 (with prob. 0.15).

(2) In the updating step the iterative structure holds at the level of a $(N+1)$-dimensional vector (conditional distribution) (see Eq. (19)). This in turn requires the computation (possibly off-line) of the transition matrix (Eq. (6)), that is, the solution of a differential equation system of dimension $N+1$.
(3) The two approximate filters feature simple iterative structures both in the prediction and updating steps. Their implementation does not require any complex calculation; thus they can be used on-line (as the well known Kalman filter) and are simple enough to be implemented in high speed networks.
(4) The approximate filters behave quite satisfactorily as compared to the exact one. Since they feature a simpler computational structure, they provide a valid interesting alternative.
(5) Processing real data (obtained form the ABILENE router) confirms the accuracy of our filters and, as a side result, the validity of the linear birth and death model assumed in Eq. (1).
(6) The analysis of the filter robustness shows that all filters are highly insensitive with respect to variations in λ, μ values. A similar robustness property holds with respect to N for the exact and Gaussian cases, while the binomial filter exhibits a significant sensitiveness with respect to N itself.
(7) Forecasting problems may find a solution via free evolution of probability distribution according to some stochastic model for traffic dynamics (such as birth-and-death model of Eq. (1)).

References

[1] T. Anjali, C. Bruni, D. Iacoviello, G. Koch, C. Scoglio, S. Vergari, Optimal filtering in traffic estimation for bandwidth brokers, in: Proceedings of the IEEE GLOBECOM'03, San Franciso, USA, December 2003, pp. 3636–3640.
[2] T. Anjali, C. Scoglio, G. Uhl, A new scheme for traffic estimation and resource allocation for bandwidth brokers, Comput. Netw. 41 (6) (2003) 761–777.
[3] T. Bonald, S. Oueslati-Boulahia, J. Roberts, IP-traffic and QoS control: towards a flow-aware architecture, in: Proceedings of the 18th World Telecommunication Congress, Paris, France, 2002.
[4] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, Wiley-Interscience, New York, 1992.
[5] V. Jacobson, Pathchar—a tool to infer characteristics of Internet paths, Presented at Mathematical Science Research Institute, April 1997.
[6] M. Jain, C. Dovrolis, End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput, in: Proceedings of the ACM SIGCOMM'02, Pittsburgh, USA, August 2002, pp. 295–308.
[7] M. Jain, C. Dovrolis, Pathload: a measurement tool for end-to-end available bandwidth, in: Proceedings of the PAM'02, Fort Collins, USA, March 2002.
[8] S. Keshav, A control-theoretic approach to flow control, in: Proceedings of the ACM SIGCOMM'91, Zurich, Switzerland, September 1991, pp. 3–15.
[9] L. Kleinrock, Queueing Systems, Wiley, New York, 1975.
[10] A. Kolarov, A. Atai, J. Hui, Application of Kalman filter in high-speed networks, in: Proceedings of the IEEE GLOBECOM'94, San Francisco, USA, November 1994, pp. 624–628.
[11] S.L. Lauritzen, Extremal Families and Systems of Sufficient Statistics, Lecture Notes in Statistics, Vol. 49, Springer, Berlin, 1998.
[12] MRTG Website. http://people.ee.eethz.ch/oetiker/webtools/mrtg/.
[13] K. Nichols, V. Jacobson, L. Zhang, A Two-bit Differentiated Services Architecture for the Internet, IETF RFC 2638, July 1999.
[14] J.W. Robets, Traffic theory and the Internet, IEEE Commun. Mag. 39 (1) (2001) 94–99.
[15] A. Terzis, L. Wang, J. Ogawa, L. Zhang, A two-tier resource management model for the Internet, in: Proceedings of the IEEE GLOBECOM'99, Rio de Janeiro, Brazil, December 1999, pp. 1779–1791.

T. Anjali received the (Integrated) M.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay, in 1998 and Ph.D. degree from Georgia Institute of Technology in May 2004. Currently, she is an Assistant Professor at the Electrical and Computer Engineering department at the Illinois Institute of Tecnology. Her research interests are to investigate Quality of Service (QoS) issues in the Next Generation Internet (NGI).

C. Bruni was born in Rome (Italy) on March 6,1939. He received the degree in Electronic Engineering from the University of Rome "La Sapienza" in 1963 and the "Libera Docenza" in Automatic Control in 1969. He became Assistant Professor in 1966, Associate Professor in 1969 and Full Professor at the University of Ancona in 1975. Since 1977, he is a Full Professor of Optimal Control at the University of Rome "La Sapienza". From 1992 to 1998 he was Director of the Research Center for Biomedical System at the same University, where, since 1992, he also is Director of the Ph.D. Course in System Engineering. His research interests are in the area of modelling, identification and data analysis, optimal control and biomedical applications. In the same fields he is author or co-author of over 80 scientific papers or books.

D. Iacoviello was born in Rome, Italy, in 1968. She received the Laurea degree in Mathematics and the Ph.D. degree in Systems Science Engineering from the University of Rome "La Sapienza" in 1992 and 1998; from 1998 to 2001 she held a postdoctoral position at the Department of Computer and Systems Science of the same University, where she is currently Assistant Professor. Her research interests are in 1D and 2D signal processing, estimation theory, optimal control and traffic estimation and resource allocation.

G. Koch was born in Rome, Italy, in 1942. He received the Laurea degree in Electrical Engineering from the University La Sapienza, Rome, in 1966, and the M.S. degree in Engineering from the University of California, Los Angeles, in 1970. He was a Full Professor in probability theory from 1976 to 1993 at School of Science, University La Sapienza. Since 1993, he has been with the University of Rome Tre. He held visiting positions with the Department of Electrical Engineering, University of California, Los Angeles, the Technische Hogeschool Twente, Enschede, The Netherlands, and the Department of Mathematics, Iowa State University, Ames. He is member of the Editorial Board of Acta Applicandae Mathematicae, the Italian Journal of Gastroenterology, and System Analysis-Modeling-Simulation. His main research interests are in linear and nonlinear stochastic dynamical systems, Martingale theory and point processes, filtering and estimation, stochastic models and identification with applications to biology and reliability.

C. Scoglio received the Dr. Ing. degree in Electronics Engineering from the University of Rome "La Sapienza", Italy, (summa cum laude) in May 1987. From June 1987 to June 2000, she had been with Fondazione Ugo Bordoni, Roma, where she was a research scientist at the TLC Network Department—Network Planning Group. In the period November 1991–August 1992 she had been visiting researcher at "Georgia Institute of Technology" "College of Computing" in Atlanta, Georgia, USA. Since September 2000, she is with the "Broadband and Wireless Networking Laboratory" of the Georgia Institute of Technology as a Research Engineer. Her research interests include network design and management.