Neural-based Queueing System Modelling for Service Quality Estimation in B-ISDN Networks

Alex Aussem†, Sébastien Rouxel, Raymond Marie‡

†*Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)*
Université Blaise Pascal, B.P. 125, 63173 AUBIERE Cedex FRANCE. e-mail : alex@sp.isima.fr
‡*Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)*
Campus de Beaulieu 35042 Rennes Cedex FRANCE. e-mail : raymond.marie@irisa.fr.

This paper addresses an original scheme based on feedforward neural networks, aimed at modelling queueing systems fed with bursty traffic. A neural network is trained to anticipate the average number of waiting cells, the cell loss rate and the coefficient of variation of the cell inter-departure time, given the mean rate, the peak rate and the coefficient of variation of the cell inter-arrival time. Our long-term goal is the design of a preventive control strategy in B-ISDN networks based on distributed neural networks modelling each queueing system located at the input and output ports of the switching facilities. To illustrate the potential of neural networks for modelling queueing systems, a neural network is successfully trained to model OnOff/D/1/c, OnOff/OnOff/1/c and multi-OnOff/D/1/c queueing systems.

Keywords – Neural networks, B-ISDN networks, traffic control, queueing systems.

1 Introduction

The future information superhighways will allow a flexible integration of a wide spectrum of services, with different traffic characteristics. These B-ISDN (Broadband Integrated Services Digital networks) based on the ATM (Asynchronous Transfer Mode) switching technique, are expected to handle multimedia services including data, voice and video traffics in a simple an unified fashion. In return, they will be subject to uncertainties about statistical variations and burstiness in arrival patterns of traffic. Thus, congestion will still be experienced due to variation in network conditions. The origin of this problem is to be found in the bursty character of the stream of data cells arriving at a buffer. Arrivals are strongly correlated and cannot be accurately modeled by a Poisson process any longer. Therefore, switching gives rise to new problems in queueing theory which cannot be solved using the mathematical tools in common use in conventional queueing theory. Hence the need for traffic control methods, such as neural networks, to adapt "intelligently" to time-varying traffic conditions [11].

Several researchers have investigated the use of neural networks (NN) to capture the complex influence of the bursty traffic on the network resources with a view to perform Call Admission Control (CAC) in a multiplexer, routing and/or bandwidth allocation [1, 3, 4, 7, 14, 15, 16, 17]. Neural-based admission control strategies usually create a decision function by learning the behaviour of an operating multiplexer conveying different types of traffic. Neural networks were also employed to predict the short-term behavior of long range dependent traffic (e.g., Web traffic) [2] using multiscale decomposition methods, and to predict the parameters of a Markov modulated Poisson process that best matches the statistical descriptors of a real bursty source [5].

In this paper, we investigate a way to model a queueing system by means of a simple feedforward NN fed with three descriptors of the aggregated input traffic, namely the mean rate, the peak rate and a burstiness measure of the cell stream, here the coefficient of variation of the cell inter-arrival time. The NN is trained to output the average number of waiting cells, $\bar{N}$, the cell loss rate, CLR, and the squared coefficient of variation of the cell inter-departure time, Cv^2_{out}. The latter may serve for the calculation of the coefficient of variation of the cell inter-arrival time of the aggregated traffic which is fed as input to the next NN along the path. In this way, the overall communication network may be modelled by means of several distributed connectionist networks located on the switches with a view to achieving global resource management.

In Section 2, the traffic descriptors are discussed. Section 3 addresses how the learning capabilities of NN's are used to model various queueing systems fed with packetized voice traffic.

2 Traffic descriptors

In a network allowing the statistical multiplexing of bursty sources, it is essential to be able to characterize the cell generation process within a call. However, the vagueness surrounding the concept of burstiness stems from the difficulty of characterizing in meaningful ways the capricious nature of cell arrivals. Traffic descriptors should ideally be: i) understandable by the user, ii) significant for resource allocation, and iii) verifiable by the network. Overall, it should be also possible to deduce an easy characterization of the aggregate traffic streams by combining the descriptors of the individual sources. With the above objective in mind, we limit ourselves to: i) the average bit rate, ii) λ, the peak rate, p, and iii) the coefficient of variation of the cell interarrival time, $Cv^2 = var(X)/E^2(X)$, where X is the length of time between the beginning of the transmission of a given cell and the beginning of transmission of the previous cell. The peak rate and the squared coefficient of variation of the inter-arrival time are relevant measures when considering cell loss [8]. Variability in cell arrivals has been connected to the queueing delays cells are subject to. The general rule is that more variability corresponds to longer delays.

Now, let class i represent a set of *independent, stationary* sources with common statistical characteristics, J is the number of classes, N_i is the number of active sources of class i. Then λ_a, p_a and Cv_a^2, the aggregate traffic descriptors are approximated by:

$$\left[\sum_{i=1}^{J} N_i \lambda_i, \quad \sum_{i=1}^{J} N_i p_i, \quad \frac{1}{\lambda_a} \sum_{i=1}^{J} N_i \lambda_i C v_i^2.\right]$$

This amounts to considering the aggregated trafic as a renewal process [9, 10]. This is a good approximation when J, the number of superimposed sources, goes to infinity.

3 Experiments

In this section, we report on experiments carried out to model a range of single server queueing systems. Several NN's are trained to model OnOff/D/1/c, OnOff/OnOff/1/c and multi-OnOff/D/1/c queueing systems. Service policy is first-in-first-out (FIFO). Such queues are ideal representations of the queueing systems encountered in telecommunication networks and switching devices [1].

3.1 Simulation & Training set-up

The data sets used for training and validation purposes were obtained by discrete-event simulation and selected such that they cover as many traffic pattterns as possible in a critical range where congestion is observed. 80% of the data was randomly selected for training and the remaining data was set aside for performance assessment. The target values, namely $\bar{N}$, CLR, and Cv_{out}^2, were estimated from the simulation. Since small CLR values (around 10^{-4}) are related to rare events, and thus subject to uncertainties, we implicitly assumed that the average CLR calculated over 10 replications was the correct answer. We used a logarithm conversion to map all parameter values in the $(0, 1)$ interval. The log transformed values were then linearly rescaled such that for each input variable X, $\bar{X} = 0$ and $Var(X) = 1$. Notice that the NN are trained to output the average logarithm, instead of the logarithm of the average, and the NMSE values given in this paper refer to squared differences of log-transformed data.

Fully-connected feed-forward NN with a single hidden layer were trained with the standard back-propagation algorithm. Momentum, weight decay and adaptive learning step were used to prevent local minima and avoid overfitting as discussed in [13]. For simplicity, each NN has 3 output neurons for the simultaneous prediction of the 3 targets. A standard measure of fit for the NN is given by the *normalized mean squared error*, that is, the mean squared error divided by the empirical variance of the simulation values calculated over all the N training patterns. A value of the NMSE = 1 thus corresponds to predicting the unconditional mean.

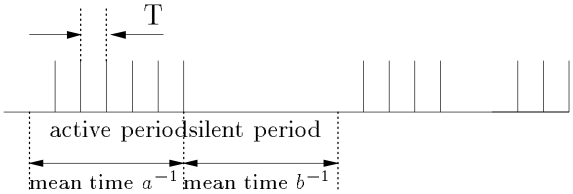

Figure 1: Traffic generated by an OnOff source

3.2 The OnOff/D/1/c queue

We consider a queue with constant service time, limited capacity ($c = 100$) and FIFO service policy but the input traffic is now generated by a single OnOff source, a simple and common traffic scenario modelling packetized *voice* source by a succession of exponentially distributed *On* and *Off* states with mean a^{-1} and b^{-1} seconds respectively (Fig. 1). In the *On* state, cells are generated every T seconds while an *Off* state corresponds to a period of silence. The descriptors of the OnOff source may be directly derived by the formulas [8, 12]:

$$\lambda = \frac{a^{-1}}{T(a^{-1}+b^{-1})},\ p = \frac{1}{T},\ Cv^2 = \frac{a(2-aT)}{(a+b)^2 T}.$$

Let $t_{on} = a^{-1}$ be the average time the source is on. The mean arrival rate, the peak rate and Cv^2 are calculated and then used as inputs to the NN. The parameters monitored are $\bar{N}$, the CLR, and Cv^2_{out}. Service rate is 13,000 cell/ms. The source parameters are randomly chosen such that, for all sources, the peak rate varies between 15,000 and 17,000 cell/ms, the average input rate varies between 6,500 and 13,000 cell/ms, and t_{on} varies between 10^{-3} and 10^{-2} ms. With these values, the CLR is found experimentally to vary between 10^{-4} and 10^{-2}.

The training set and the test set are made up from 456 examples and 75 examples respectively. All values were rescaled as discussed above. Best results were obtained with 15 hidden units. The NN outputs on the test set are plotted in Fig. 2. As it may be seen, the prediction of $\bar{N}$ and Cv^2_{out} are very accurate. The NMSE on the test set for $\bar{N}$ is $5.74 \cdot 10^{-4}$, $5.27 \cdot 10^{-5}$, for Cv^2_{out} and $1.36 \cdot 10^{-3}$ for CLR. Prediction accuracy degrades somewhat for small CLR values owing to the small number of training examples with small CLR's and the statistical uncertainty of these values.

3.3 The OnOff/OnOff/1/c queue

In this section the server is OnOff, i.e., the server is inhibited during an exponentially distributed time and the service time is deterministic in the *On* state. The underlying idea is to consider a single deterministic server sharing multiple buffers in which transitions occur randomly according to the priority of the incoming cells in the distinct buffers. After serving a cell, the server switches to the *Off* state with probability p^s. It remains in this state for T^s, the constant service period. Activation and sleeping periods follow a geometric law with parameter p^s and q^s respectively. The traffic is generated by a single OnOff source. In such systems, the difficulty lies in the strong dependence of CLR to the input traffic as well as the state of the server. In the active state, the server can only switch to another state at the end of a service time. Therefore, the input rate strongly influences the effective period of time the server is active.

The average service rate is 13,000 cell/ms, $p^s = 0.7$ and $q^s = 0.4$, the peak rate varies between 13,000 and 20,000 cell/ms, the average input rate varies between 6,500 and 13,000 cell/ms, t_{on} varies between 10^{-4} and 10^{-2} ms. Data were rescaled as discussed before. Only the examples with a CLR in the range $[10^{-4}, 10^{-2}]$ were selected for training. A 3/15/3 feedforward NN was used as is the OnOff/D/1/c case.

Forecasts on the test set are presented in Fig. 3 on the left side. Here again, the forecasts of $\bar{N}$ and Cv^2_{out} are very accurate. The CLR prediction is slightly more difficult. The NMSE on the test set for $\bar{N}$ is $3.55 \cdot 10^{-3}$, $1.21 \cdot 10^{-4}$, for Cv^2_{out} and $8.61 \cdot 10^{-3}$ for CLR. Prediction accuracy degrades somewhat for small CLR. Still, remarkably positive statements can be made in regard to the queue load, the loss rate and the output stream burstiness.

3.4 The Multi-OnOff/D/1/c queue

In this section, several independent OnOff sources are multiplexed. For single-bit rate multiplexing, i.e., calls have the same traffic characteristics, the aggregated traffic descriptors are considered to be sufficient information about the traffic. Suppose the sources are identical with arrival rate λ, and peak rate p, satisfying

$$N\lambda \le \mu \le Np$$

The number and the characteristics of the OnOff sources were randomly varied, $N \in [10, 100], \mu = 13000$ cells/ms and λ randomly varied such that the above inequality holds. For each traffic pattern, CLR, was obtained upon simulation and only the patterns with $10^{-4} < CLR < 10^{-2}$ were selected for training.

Since all OnOff sources were identical, the weighted sum on the left side of Eq.() is exactly Cv^2, the individual coefficient of variation of the cell inter-arrival time for each source. The latter is fed as input to the NN. So the aggregated traffic variability is approximated by the variability of the marginal traffics. This variable is independent upon the number of sources that are multiplexed. The mean rate and the peak rate, however, are dependent on the number of sources.

A 3/15/3 NN was trained using with 533 examples, and validated over 55 patterns. 10,000 training epochs were necessary. Results are shown in Fig. 3, right side. The NMSE values on the test set are $1.28 \cdot 10^{-2}$ for $\bar{N}$, and $2.71 \cdot 10^{-4}$ for Cv_{out}^2 and $1.05 \cdot 10^{-2}$ for CLR. Results are not as good (as expected) but still acceptable. The integer part of $\log(CLR)$ is correct. Cv_{out}^2 is fairly well forecast, this is very encouraging since the latter should serve as input to the next NN on the path.

3.5 Discussion

The preliminary results presented here are useful in proving that the NN approach is effective in simple settings. It is our hope that more realistic traffics and queueing systems could be modelled too in order to provide a baseline for general resource management. Since each NN was found to provide a reliable estimate of the output traffic burstiness, Cv_{out}^2, could be transmitted to the NN modelling the next queueing system on the route and so on. Therefore, a complete communication path could be modelled by means of several NN distributed on the queueing systems. Given $\bar{N}$ at each queue and using the Litlle formula [9], the end-to-end cell delay could be estimated as well as the total cell loss rate. Experiments on neural-based trafic control strategies on tandem queues are being conducted at present and will be reported in due course.

4 Conclusion

Satisfactorily tested on several queueing systems, the neural-based approach was shown to adapt to different packetized voice traffic conditions without structural modifications. Very accurate estimations are obtained for the average number of waiting cells, the cell loss rate and the outgoing traffic burstiness in all cases. This is encouraging since burstiness measures may be transmitted from queue to queue to allow for global queueing network traffic management. This is left for future work.

References

[1] A. Aussem, "Call admission control in ATM networks with the random neural network," *Proceedings of the IEEE Int. Conf. on Neural Networks*, Vol. 4, pp. 2482-2487, 1994.

[2] A. Aussem and F. Murtagh, "A neuro-wavelet strategy for Web traffic forecasting," *Research in Official Statistics*, Vol. 1, No. 1, pp. 65-87, 1998.

[3] R. Bolla, F. Davoli, P. Maryni, T. Parsini, "An adaptive neural network admission controller for dynamic bandwidth allocation," *IEEE Trans. Syst., Man, and Cyb.*, Vol. 28, No. 4, 1998.

[4] H. Brandt, E. Nordström, O. Gällmo and M. Gustafsson and L. Asplund, "A hybrid neural network approach to ATM admission control", *International switching symposium World telecommunications Congress, ISS'95*, Berlin, pp. 283-287, 1995.

[5] E. Casilari, A. Alfaro, A. reyes, A. Diaz-Estrella, F. Sandoval, "Neural modelling of Ethernet traffic over ATM networks", *EANN'98*, June, Gibraltar, pp. 304-307, 1998.

[6] D.R. Cox and H.D. Miller, *The theory of Stochastic Processes*, London, Methuen, 1965.

[7] X. Chien and I.M. Leslie, "A neural network approach towards adaptive congestion control in broadband ATM networks," *Proc. IEEE Globecom'91*, pp. 115-119, 1991.

[8] COST 242, “Methods for the Performance Evaluation and Design of Broadband Multiservice Networks,” Parts I Traffic Control, Final Report, Seminar version, June, 1996.

[9] E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks*, 2nd Ed., Wiley, 1998.

[10] R. Gusella, “Characterizing the variability of arrival processes with indexes of dispersion” *IEEE J. Sel. Areas Commun.*, No.2, Vol.9, pp. 203-211, 1991.

[11] I.W. Habib, “Applications of neurocomputing in traffic management of ATM networks”, *Proceedings of the IEEE*, No. 10, Vol. 84, Pp. 1430–1441, 1996.

[12] H. Heffes, D.M. Lucantoni, “Markov Modulated characterization of packetized voice and data traffic and related statistical multiplexer performance,” *IEEE J. Sel. Areas Commun.*, Vol. SAC-4, No 6, 1986, pp. 856-867.

[13] J. Hertz, A. Krogh et R. Palmer, *An Introduction to the Theory of Neural Computation.* Redwood City, CA: Addison-Wesley, 1991.

[14] A. Hiramatsu, “Integration of ATM Call Admission Control and Link Capacity Control by distributed neural networks,” *IEEE J. Sel. Areas Commun.*, Vol. 9, 1991, pp. 1131-1138.

[15] J.E. Neves, L.B. Almeida, and M.J. Leitco, “B-ISDN connection admission and routing strategy with traffic prediction by neural networks,” *Proceedings of the International Conference on Communications SUPERCOM - ICC’94*, New Orleans, Louisiana, pp. 769-773, 1994.

[16] E. Nordström, J. Carlström, O. Gällmo and L. Asplund, “Neural networks for adaptive traffic control in ATM networks,” *IEEE Communications Magazine*, Oct. 1995.

[17] P. Tran-Gia and O. Gropp, “Performance of a neural net used as admission controller in ATM systems,” *Proc. IEEE GLOBECOM’92*, Orlando, Florida, pp. 1303-1309, 1992.

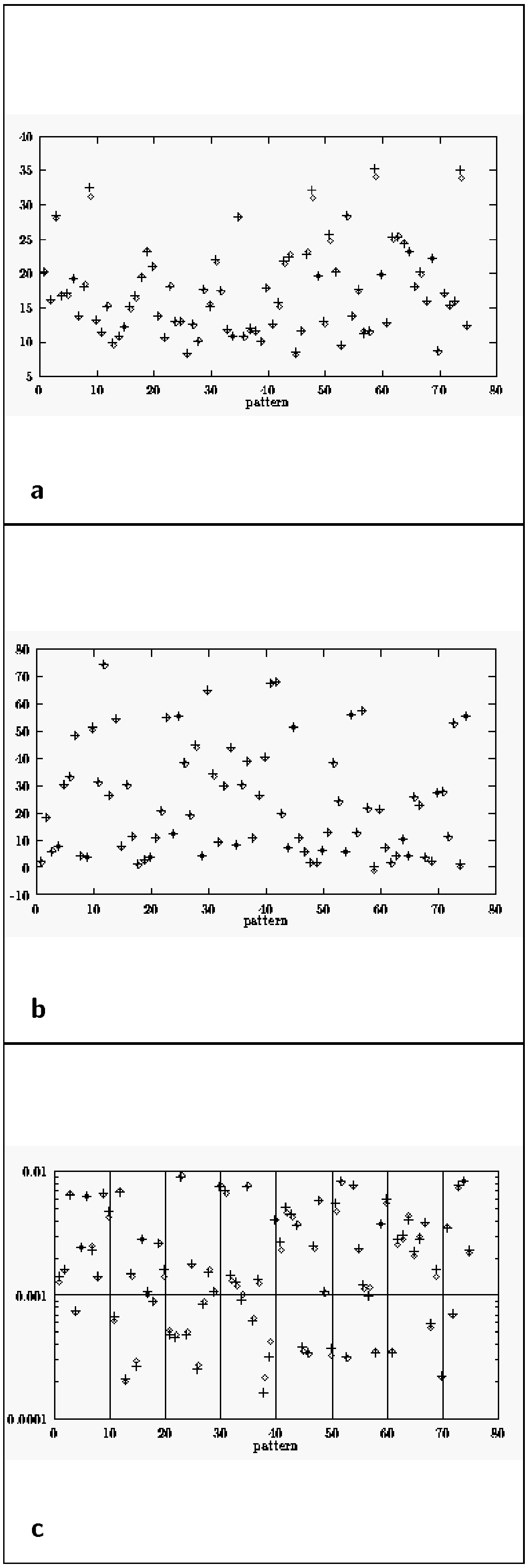

Figure 2: OnOff/D/1/c queue. $c = 100$. From top to bottom: $\bar{N}$, Cv^2_{out} and CLR. Actual value is $+$, forecast is $\circ$.

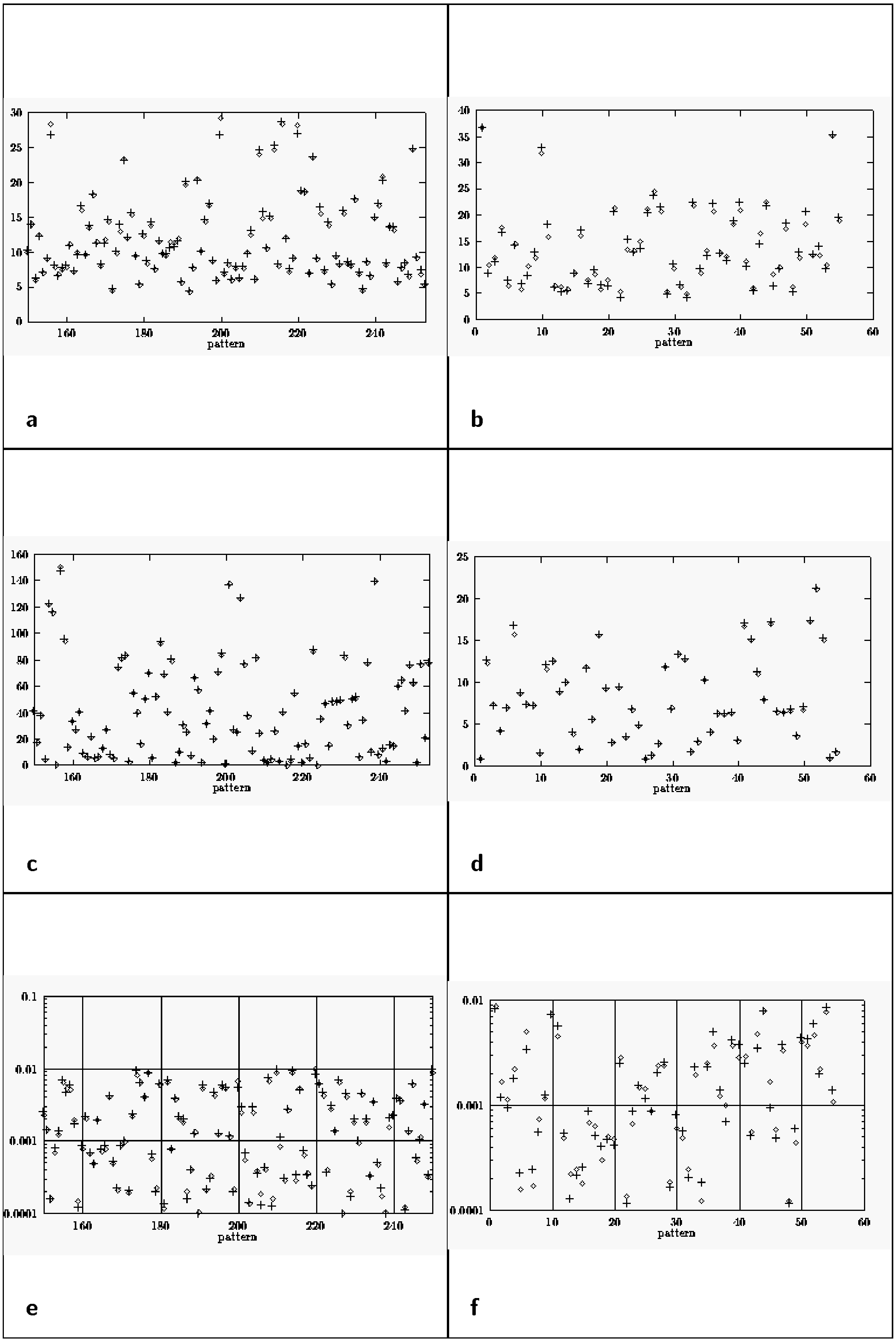

Figure 3: Left plots: OnOff/OnOff/1/c queue; right plots: N-OnOff/OnOff/1/c queue. $c = 100$. From top to bottom: $\bar{N}$, Cv^2_{out} and CLR. Actual value is $+$, forecast is $\circ$.