Nonlinear Analysis of Network Traffic Measurements

P. Akritas[1], P.G. Akishin[2], I. Antoniou[1,3], A.Yu. Bonushkina[2], I. Drossinos[4], V.V. Ivanov[1,2], Yu.L. Kalinovsky[2], V.V. Korenkov[2] and P.V. Zrelov[2]

[1]*International Solvay Institutes for Physics and Chemistry, CP-231, ULB, Bd. du Triomphe, 1050, Brussels, BELGIUM*

[2]*Laboratory of Information Technologies, Joint Institute for Nuclear Research, 141980, Dubna, RUSSIA*

[3]*Department of Mathematics, Aristotle University of Thessaloniki, 54006 Thessaloniki, GREECE*

[4]*European Commission, Joint Research Center, I-21020 Ispra (Va), ITALY*

We applied a nonlinear time series analysis approach to the traffic measurements obtained at the input of a medium size local area network. We used data at the input of the Dubna University network, including about 200 interconnected computers. The performance of the data acquisition system was realized on the basis of standart IBM PC: Fig. 1.

The recorded traffic data correspond approximately to 20 hours of measurements. The part of this series corresponding to 1 hour of measurements is presented in Fig. 2.

Estimation of dynamical system parameters

The key idea of this approach is that the correct space for viewing the structure of the dynamical system that generates the signal is not the one-dimensional space of observables $\{x_i\}$. The dynamics takes place in a space of vectors $\vec{y}_i$ of larger dimension:

$$\vec{y}_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}), \tag{1}$$

Here m is the dimension of the dynamical system, and τ is a "delay time", or the correlation length of series $\{x_i\}$.

To choose the independent components from the traffic data, we computed the correlation length τ. It corresponds to the point, where the linear auto-correlation function first crosses the confidence tube corre-

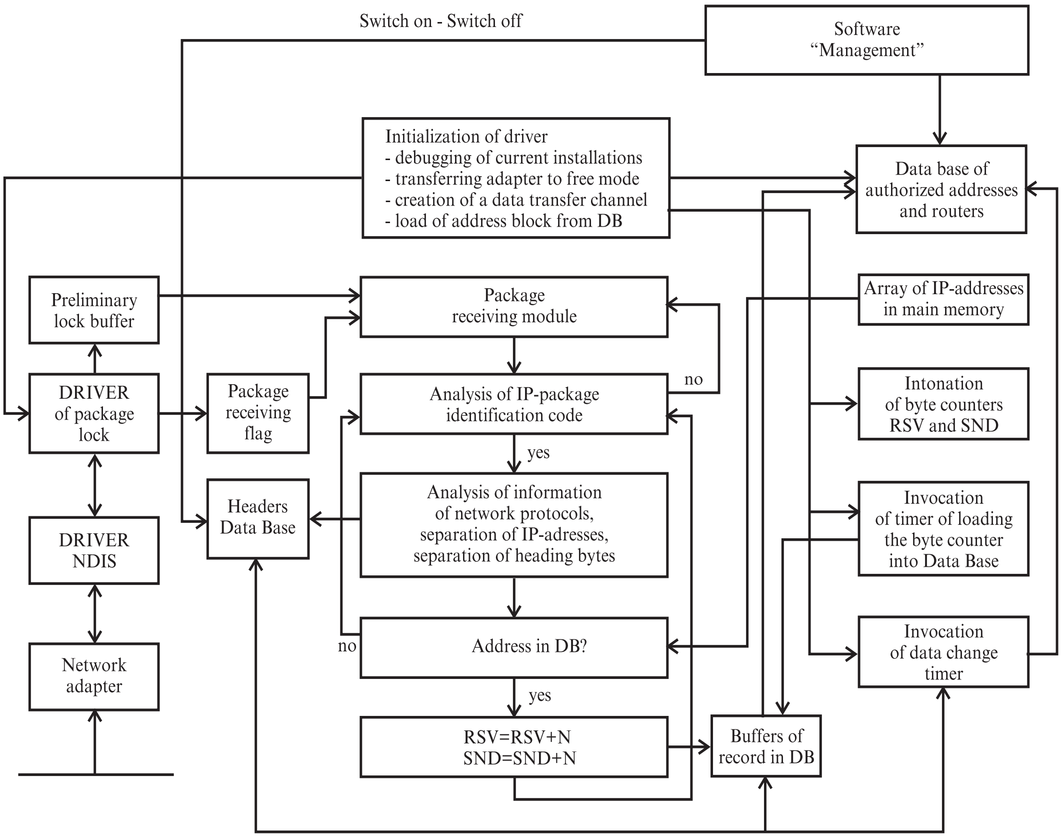

Figure 1: The scheme of data acquisition system

sponding to the Gaussian white noise.

$$C(\tau) = \frac{\sum\limits_{i=1}^{N}(x_{i+\tau} - \bar{x})(x_i - \bar{x})}{\sum\limits_{i=1}^{N}(x_i - \bar{x})^2} \tag{2}$$

The correlation length against the aggregation bin size is plotted in Fig. 3.

We can see that for bin sizes from 0.1 sec up to 10 sec the correlation length τ is ~ 10 sec. The points separated by the time interval τ can be considered as linear independent.

The embedding dimension counting for aggregated time series has been performed with the Grassberger-Procaccia algorithm:

$$C_n^m(r) = \frac{2}{N(N-1)} \sum_{i \neq j} \Theta(r - |\mathbf{y_i} - \mathbf{y_j}|), \tag{3}$$

with the distance between two points given by

$$|\mathbf{y}_i - \mathbf{y}_j| = \max\left\{|x_i - x_j|, ..., |x_{i+(m-1)\tau} - x_{j+(m-1)\tau}|\right\}.$$

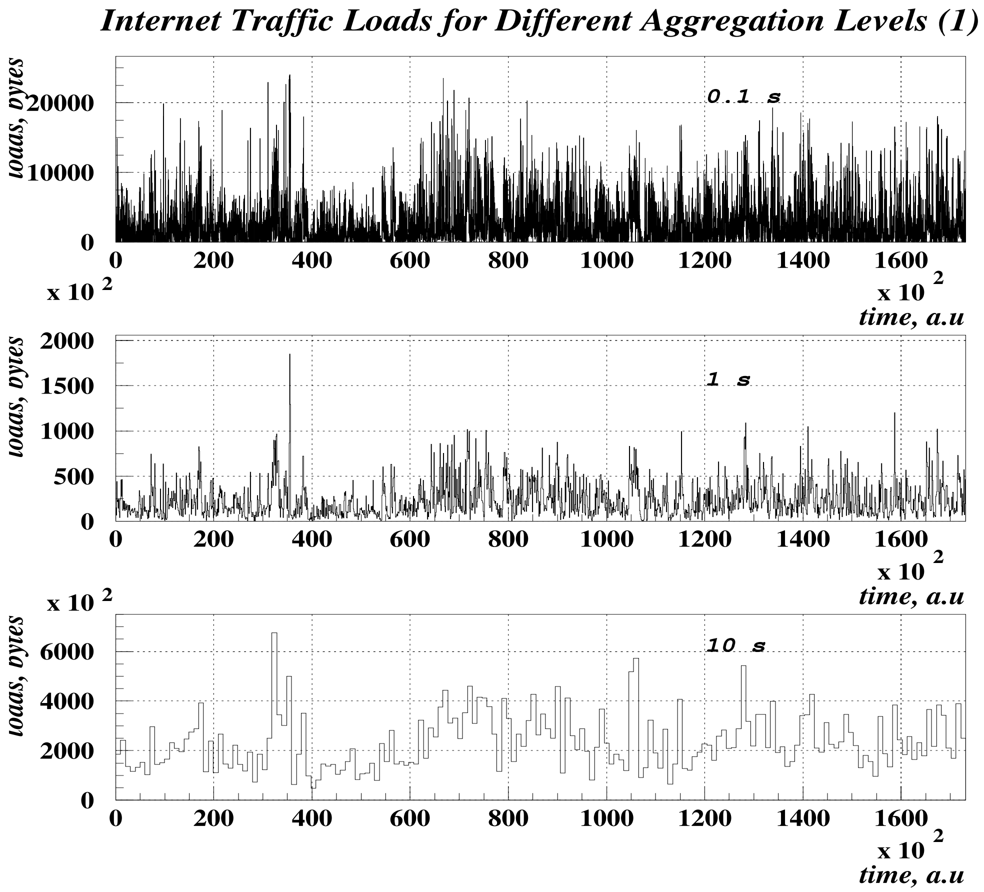

Figure 2: Traffic measurements aggregated with different bin sizes: 0.1 sec, 1 sec, 10 sec

Here $\Theta = 1$ if its argument is non-negative and 0 otherwise.

The value of the correlation integral $C_n^m(r)$ is the empirical probability that a randomly chosen pair of points will be separated by a distance less then $\leq r$. The integral $C_n^m(r)$ is computed for r ranging from 0 to the largest possible value and for m increasing from 1 up to the largest possible value. The slope of the logarithm $\log C_n^m(r)$ versus $\log r$ gives the lowest estimate of the embedding dimension: Fig. 4.

For different parts of the time series analysed, no saturation of the slope with respect to increasing m was found. According to the Takens theorem this may imply a very high dimension of the studied time series.

Wavelet filtering

The traffic measurements can be presented as the sum of a regular process and a stochastic part, related to the high frequency. Elimination

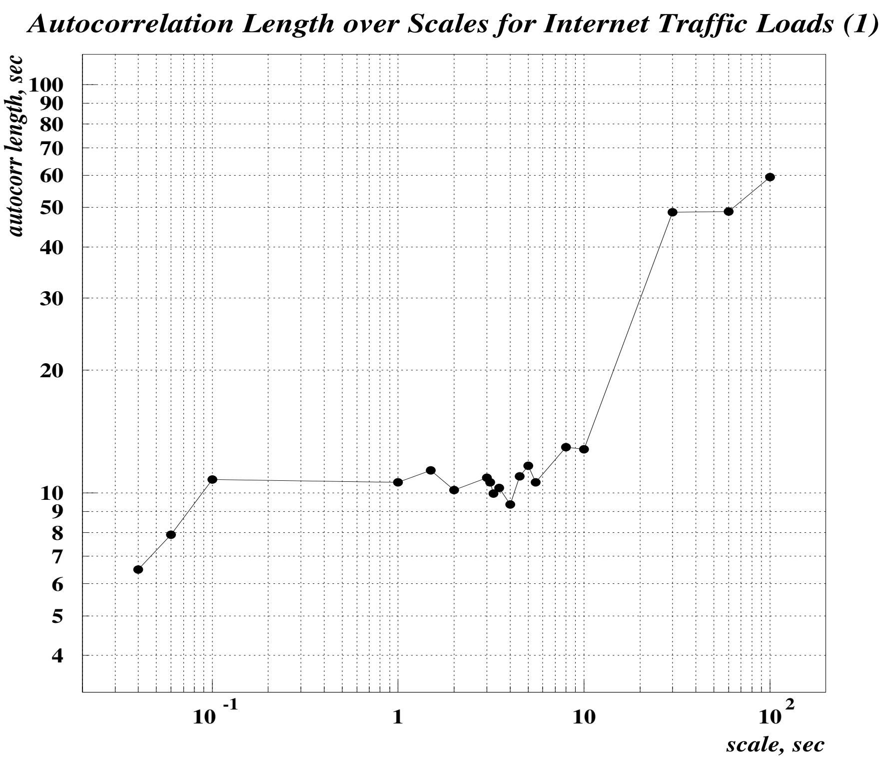

Figure 3: The dependence of the correlation length against the size of the aggregation bin

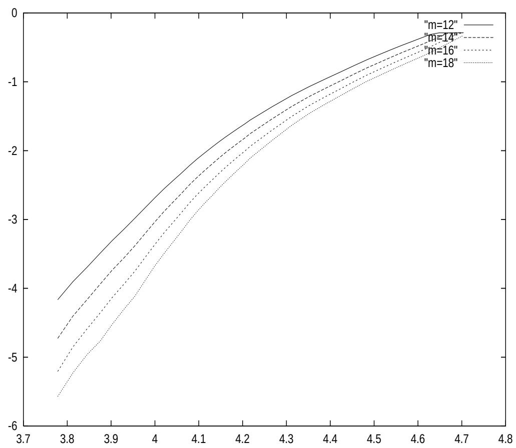

Figure 4: Correlation integrals $C_2^m(r)$ for traffic measurements aggregated with 1 sec bin: $\tau = 10$ sec and m =12, 14, 16, 18

of the *noisy* part may simplify the analysed data and reduce the dimension of the dynamical process. In order to do this, we used the discrete wavelet transform based on a class of the Daubechies wavelets. In Figure 5 we present a part of the original traffic series, the corresponding filtered signal and the noisy component.

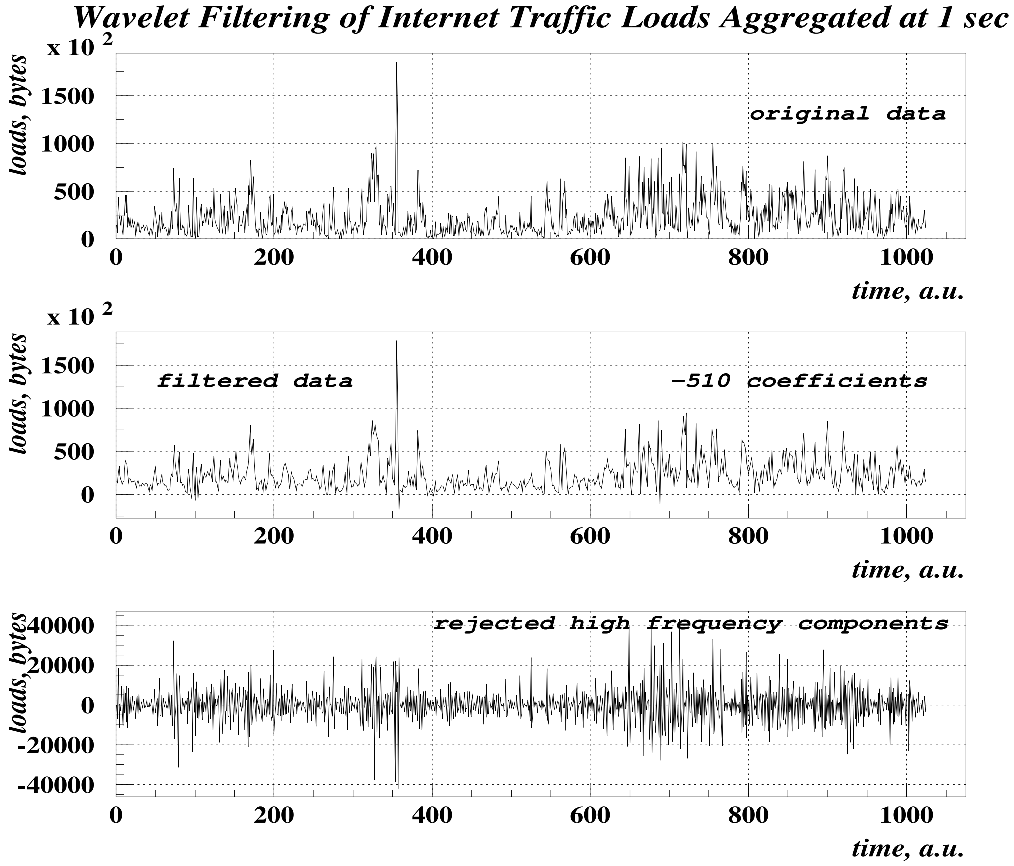

Figure 5: Traffic measurements: a) original traffic series, (b) filtered signal, (c) noisy component

After filtering out the noisy component from traffic measurements, we estimated the Grassberger-Procaccia dimension for the remainders. What was observed for all curves is that the dimension about $16 \div 18$ seems to be close to saturation.

Principal components analysis of traffic data

An additional method for the clarification of the dimensionality problem was the Principal Components Analysis. We applied this method to the original traffic data and to the filtered data. Figure 6 shows the PCA results. It can be seen that the intrinsic information of the analyzed system is accumulated in the few first components.

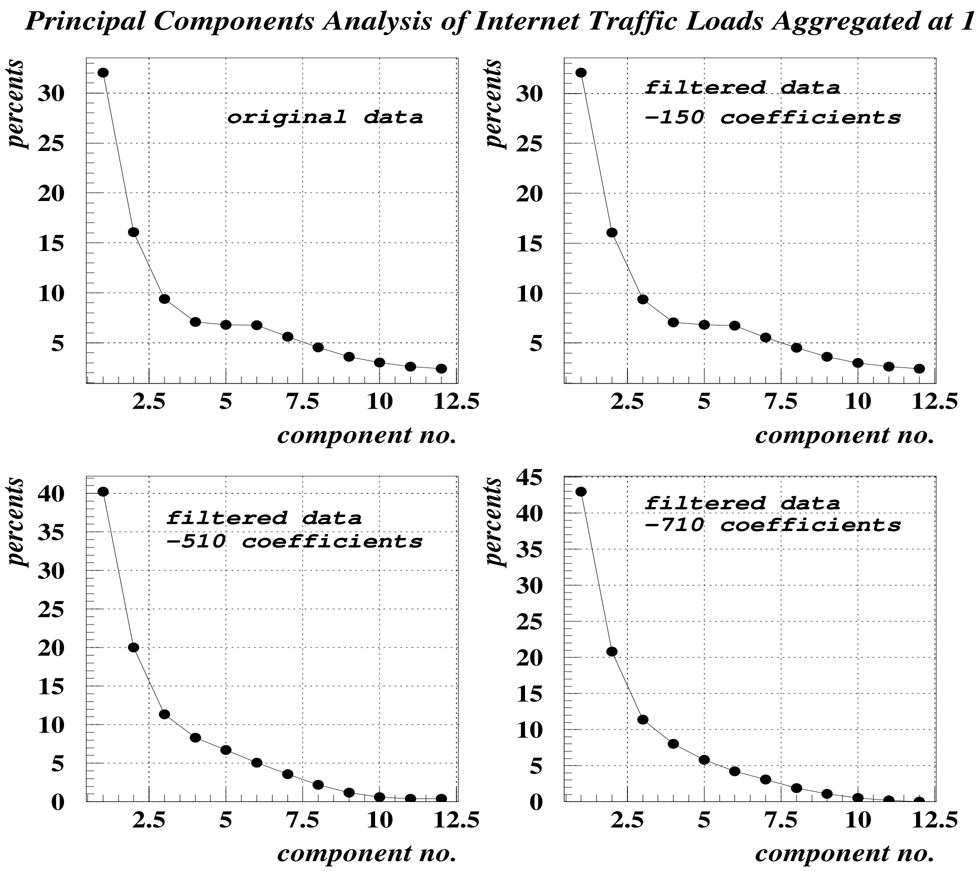

Figure 6: Principal components: (a) for original data, (b)-(d) for time series filtered at different thresholding level

Reconstruction of underlying dynamical system

To reconstruct the dynamical system corresponding to the traffic measurements, we used an artificial neural network (ANN). In our study we applied a layered neural network with the feed-forward architecture from the JETNET3 package. For the ANN training we used a data set corresponding to approximately 34 minutes of measurements and aggregated with time bin 1 *sec*. These data were first cleaned applying wavelet filtering for elimination of "noisy" component. The following parameters were used for the input vector formation: $\tau = 10\,sec$ and $d_E = 15 \div 20$. Figure 7 presents a part of the traffic data (traffic.dat) and the result of the ANN approximation (train.dat) after 1000 training epochs.

Figure 8 shows the statistical distributions of *Errors* and *Relative Er-*

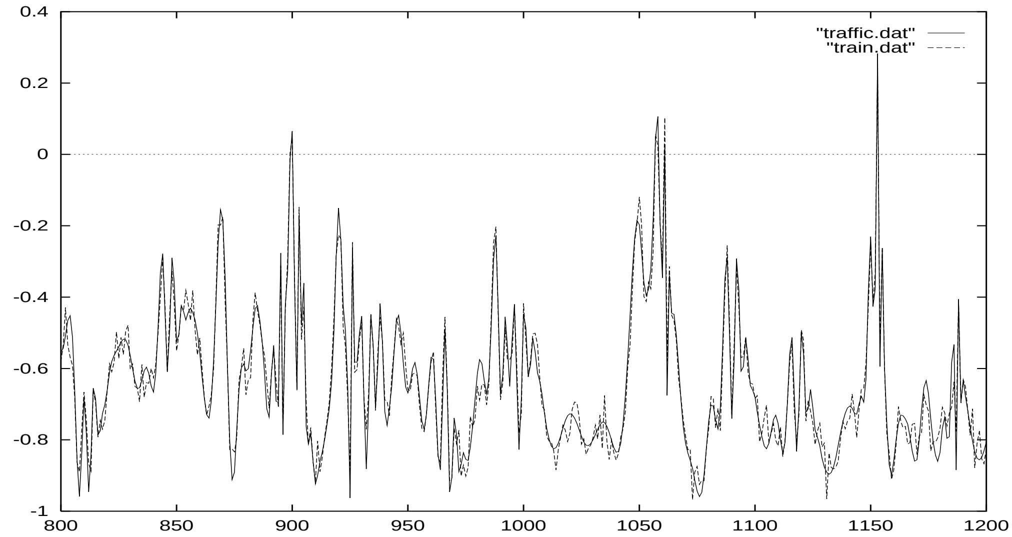

Figure 7: The result of the ANN approximation of the traffic series after 1000 training epochs

rors after 1000 training epochs. We can see that the ANN demonstrates a high level of data approximation.

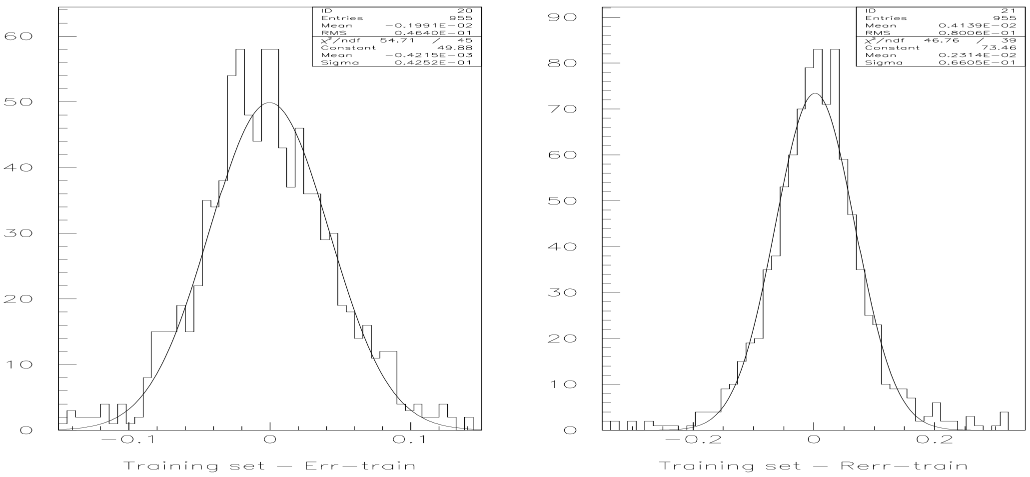

Figure 8: Distribution of *Errors* (left) and *Relative Errors* (right)

Nonlinear model of network traffic based on ANN

The ANN after training can be used as a model for the simulation of traffic measurements. It can be seen as a nonparametric model, because this approach does not require any assumptions concerning possible parameters of the analysed time series.

Figure 9 demonstrates the distributions of sizes of traffic packages (normalized to the interval [-1,1]) for the original traffic measurements and for time series generated by the ANN. We see that the neural network model reproduces the statistical distribution of real data quite well.

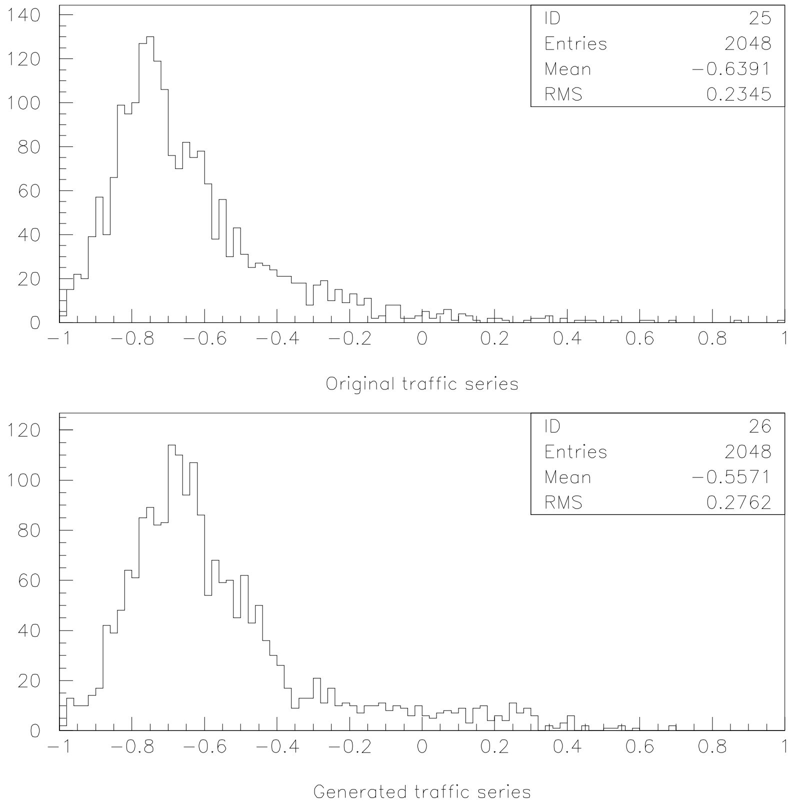

Figure 9: The distribution of sizes of the traffic packages (normalized to the interval [-1,1]) for : a) the original traffic measurements, and b) the generated by the trained ANN

Conclusion

In our work we systimatically applied the nonlinear time series analysis to the traffic measurements obtained at the input of the medium size local area network. Using the estimated values for the time lag and the embedding dimension we successfully applied a layered feed-forward neural network for the identification and reconstruction of the dynamical system underlying the traffic measurements. We found that the trained neural network reproduces the statistical features of real traffic data.