A NEW APPROACH TO THE MODELING OF NETWORK TRAFFIC IN SIMULATIONS

Matjaž Fras, Jože Mohorko, Žarko Čučej

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija

Key words: self-similar, network traffic modeling, Pareto distribution, maximal transmission unit

Abstract: Simulations of telecommunications networks have become very important tools for their evaluation. A very important influence in simulations has network traffic. This paper introduces new concepts for the modeling of measured network traffic in simulation tools. With these new concepts, we can improve descriptions of the random packet-size process, especially for maximal-packets of network traffic, which have a very great impact on the bit or packet rates of network traffic. The suggested methods improve the contents of packets, especially maximal packets in modeled network traffic simulations, which leads to smaller differences in bit and packet-rates between measured and modeled network traffics.

Nov pristop k modeliranju samo-podobnega prometa v simulacijah

Kjučne besede: samo-podobnost, modeliranje omrežnega prometa, Pareto porazdelitev, maksimalna dolžina paketa

Izvleček: Simulacije telekomunikacijskih omrežij postajajo pomembno orodje za ovrednotenje le teh. Zelo pomemben in velik vpliv v simulacijah ima tudi omrežni promet. Ta članek predstavlja novi koncept modeliranja izmerjenega omrežnega prometa v simulacijskih orodjih. Z tem novim konceptom lahko izboljšamo opis naključnega procesa velikosti paketov omrežnega prometa, zlasti maksimalnih paketov, kateri imajo zelo velik vpliv na srednjo vrednost celotnega prometa v bitih in paketih na časovno enoto. Predlagane metode izboljšajo opis vsebnosti paketov v omrežnem prometu, še posebej maksimalnih paketov v modeliranem prometu,kar posledično vodi do manjših razlik v srednji vrednosti bitov in paketov na časovno enoto med izmerjenim in modeliranim prometom.

1. Introduction

Statistical analysis in Ethernet networks show that, in many cases, network traffic can be described by self-similarity /1/. This model appeared before fifteen years as an alternative, at that time, to the used models such as Poisson and Markov /2/. It was also shown, that heavy-tailed distributions, such as Pareto and Weibull, are more suitable for describing network processes, such as process packet-size and inter-arrival time /1, 3, 4, 5/.

One of the main goals of researchers was, and still is, the modeling of network traffic in simulations, such as OPNET /6, 7, 8/. In simulation we try to model the measured network traffic, which is the best possible approximation of the measured traffic in the sense of bit or packet-rates, bursts or variance. For evaluating discrepancies between measured and simulated network traffic, we chose different measures such as bit or packet rates, Hurst parameter, variance and also discrepancy between histograms of statistical network process for packet size and inter-arrival time.

During measuring and modeling we saw that discrepancies between measured and modeled traffic are derived from an inaccurate description of the packet-size process. We also saw that, especially for longer and maximal length packets (MTU- Maximal Transmission Unit), have a substantial influence on modeled network traffic. The captured histogram of the packet-size process had great discrepancy in regard to the measured histogram and chosen distribution, which is usually a consequence of maximal-packets. Maximal packets are a consequence of data fragmentation in TCP/IP stack. Usually with classical modeling, where a captured histogram of packet size process is described with distribution, we do not derive at a good enough description regarding the packet-size process of measured traffic, especially the content of maximal packets. This, consequently, leads to great discrepancy between measured and modeled network traffics, especially in bit and packet-rates, and also traffic bursts.

For this reason, we present three methods for describing a measured traffic histogram of packet-size which achieve more accurate descriptions of network traffic in simulations.

1. The first method is based on using "mixed distributions" for describing random processes, a similar concept is used in the area of image processing /9/, and already steps in the area of traffic modeling /10, 11, 12/.
2. The second method is based on estimating data files of a measured traffic histogram by defragmentation in a communications network /3, 4, 5/.
3. The third method combines the first and second methods.

This paper is organized as follows. The second section describes statistical modeling of network traffic by distribution and Hurst parameter. The next section describes

the packet-size process of network traffic. New approaches with suggested methods are in the forth section. The fifth section represents the simulation results. Finally, we finish this paper with the conclusion.

2 Statistical modeling of measured network traffic

Network traffic can be described as a combination of two random processes:

1. packet-size process $X(t)$
2. inter-arrival time $Y(t)$

Lets describe network traffic Z(t) as

$$Z(t) = \psi(X(t), Y(t)) \quad (1)$$

where ψ is the function of packet-size $X(t)$ and inter-arrival time process $Y(t)$. Both processes are described by probability distribution function (pdf). The choice of suitable distribution for a traffic process depend the measured network traffic's properties. For network traffic with a short-range dependence property, light-tailed distributions (exponential) are the more suitable for describing packet-size process, such as exponential. In the case of network traffic with long-range dependence, heavy tailed distributions are the more suitable distributions for describing such traffic, such as Pareto and Weibull. The probability density function (pdf) of Pareto distribution is

$$p(x) = \alpha k^{\alpha} \cdot x^{-\alpha-1}, \quad k \le x, \quad \alpha, k > 0 \quad (2)$$

where k is local parameter and α is shape parameter. Probability density function of Weibull distribution is:

$$p(x) = \frac{\alpha}{k} \cdot \left(\frac{x}{k}\right)^{\alpha-1} \cdot e^{-\left(\frac{x}{k}\right)}, \quad x \ge 0, \quad \alpha, k > 0 \quad (3)$$

where k is local parameter and α is shape parameter.

Definition of the self-similar random process /3, 13, 14, 15/ is based on autocorrelation function $r(k)$, which is described as

$$r(k) \approx k^{-\beta} L_1(k), \quad k \to \infty, \quad 0 < \beta < 1, \quad (4)$$

where $L_1(k)$ is slowly varying at infinity, that is for all $x > 0$ (i.e., $L_1(t)$ = constant, $L_1(t) = log(t)$). Hurst parameter H is used for described arrival process and it is defined by

$$H = 1 - \frac{\beta}{2}, \qquad 0 < \beta < 1 \quad (5)$$

and presents the measure of self-similarity. For describing arrival process, beside parameter H, are also needed parameters such are average arrival-rate, fractal onset time scale, source activity-ratio, and peak to mean ratio.

3 Problem of statistical packet size process

From measured traffic by sniffer /8/, we can obtain information about a packet-sizes, inter-arrival time, packet-rate... Based on histograms, we can evaluate both random traffic process $X(t)$ in $Y(t)$ and choose distributions, which are the best approximations of histograms. During research, where we estimate parameters of traffic processes we found that, in the case of estimating packet-size process parameters much larger discrepancies appear than in the case of inter-arrival time. Discrepancy between the histogram of measured traffic and distribution, which describe this process, can be evaluated by goodness of fit tests, such as Kolmogorov-Smirnov or Chi-square /16/. The greatest impact on these discrepancies is MTU, which as mentioned in the first section. MTU packets cause a strong discontinuity in the histogram and it is very difficult to describe such a histogram using the classical method. In our research, we paid attention to a statistical description the packet-size process of network traffic.

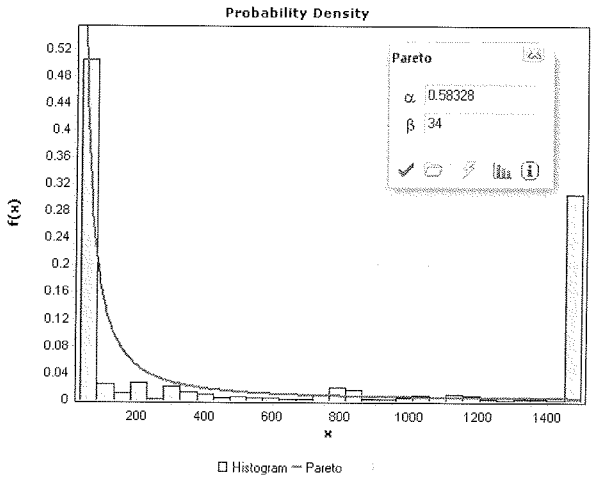

Fig. 1: Histogram of measured packet size process and distribution parameters estimation with classical method with EasyFit fitting tool.

Figure 1 shows an example of a packet-size histogram of measured network traffic and classical distribution parameters' estimation. From the captured histogram, we can see that minimal length size packets of around 54 B prevail. But there are also a lot of packets of maximal length, which also have a great influence on the bit-rate of the entire network traffic. The classical parameter estimation method (Figure 1) does not describe the process very well, especially those maximal packets, which usually lead to great discrepancies between measured and modeled traffic, in the sense of bit or packet rates. Such an estimation method also has very big difference between the contents of packets between measured and simulated traffic. We cannot solve this problem by using other methods for estimating distribution parameters for the packet-size process of network traffic, such as the CCDF method /3/.

The greatest discrepancies appear when describing network traffic with long-range dependence (LRD) property, where heavy tailed-distribution is used, such as Pareto. Smaller discrepancies also appear in the case of describing network traffic with short-range dependence (SRD), where exponential distribution was used, but these discrepancies are smaller than in the previous case.

3 Suggested methods for estimating distribution for packet-size process

All suggested methods are based on the transformation of captured-traffic. The first method is based on using mixed (multiple) distributions to the describe packet-size process, the second method is based on defragmentation of captured-packets and the third method is a combination of the first and second methods.

3.1 Mixed distributions

Using this method, we will describe network-traffic by multiple-distributions, which will be implemented using multiple traffic generators in the same simulation workstation. By using mixed distributions for describing the stochastic process of network traffic, we will achieve a smaller discrepancy between the measured histogram and the fitted distributions for packet-size process (Figure 2). Network traffic $Z(t)$ defined in (1) can be described as the sum or n-th data sources:

$$\begin{aligned} &Z(t) = Z_1(t) + Z_2(t), \ldots Z_n(t) \\ &Z(t) = \psi_1(X_1(t), Y_1(t)) + \ldots + \psi_n(X_n(t), Y_n(t)) = \\ &Z(t) = \sum_1^n Z_i(t) = \sum_1^n \psi_i(X_i(t), Y_i(t)) \end{aligned} \quad (6)$$

where $Z_i(t)$ is traffic for each traffic generator and ψ_i is a function of two random processes $X_i(t)$ and $Y_i(t)$, where $X_i(t)$ represent packet-size process and $Y_i(t)$ inter-arrival time. So, we can divide network traffic into separated segments modeled by different distributions. Points which separate the packet size process in multiple parties described by independent distribution, are threshold points. The simplest way to separate network-traffic for mixture distribution is to define the first traffic $Z_1(t)$, where are packets, which are longer than the threshold value, and another traffic $Z_2(t)$, with packets that are shorter than the threshold. In many cases, MTU size represents the threshold point.

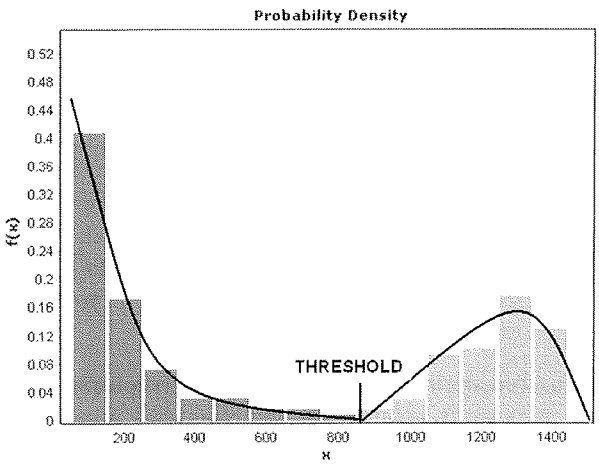

Fig. 2: Example of using two distributions for describing packet size process of captured network traffic.

$$\begin{aligned} &Z(t) = Z_1(t) + Z_2(t) = \\ &= \begin{cases} Z_1(t) = \psi_1(X_1(t), Y_1(t)); & packet\_size > threshold \\ Z_2(t) = \psi_2(X_2(t), Y_2(t)); & packet\_size \le threshold \end{cases} \end{aligned} \quad (7)$$

We must also estimate the belonging distributions for both inter-arrival time processes $Y_1(t)$ and $Y_2(t)$ and packet-size processes $X_1(t)$ and $X_2(t)$.

3.2 Defragmentation method

Whilst transmitting files across a network, IP packets are fragmented because of MTU limitations. The fragmentation process is executed in a model of IP encapsulation in TCP/IP stack. From the captured traffic in Figure 1, we can see that MTU packets impact on the discontinuity in the histogram, this causing the common distribution descriptions, with the help of the classical method. This new method is based on histogram estimation of the transmitted data file before fragmentation /4/. For a distribution estimation of the packet-size process we execute with the addition of maximal packets, which are fragmented in the fragmentation process during transmission. So, we combine all packets from a sequence of MTU packets, including the first packet shorter than the maximal size, from the same source in the new bigger packet. These newly derived at values, together with captured non-fragmented packets, are used designating the histogram of data, which will be described by new distribution.

$$\begin{aligned} &Z(t) = \psi(X(t), Y(t)) \rightarrow Z_T(t) = \psi(X_T(t), Y_T(t)) \\ &Z(t) \approx Z_f(t) \end{aligned} \quad (8)$$

$Z_T(t)$ represents the transformed traffic, which is a function of the transformed processes for packet-size X_T and inter-arrival time Y_T. The transformed histogram represents the originally transmitted files $Z_f(t)$. We spray the distribution of maximal packets in the captured histogram over a new range, using the defragmentation method, which represents the transmitted files. This method leads to more continuous histograms, such as in Figure 3, which can be described by the classical method more precisely using distribution, than the histogram in Figure 1. Estimation parameters of file sizes are used in traffic generators during simulations. Because of the limitation of MTU, which is a defined in model of a communication device, the files are fragmented into maximal packets during the simulation run. So estimate traffic is a good approximation of captured traffic.

3.3 Combination of distributions and defragmentation

The third method is based on a combination of mixed distributions and defragmentation methods. The basic idea is to describe captured traffic with two or more distributions, but for captured-traffic we can also execute the defragmentation process. For example, we can execute the defragmentation process of captured-traffic $Z(t)$, and then describe with one distribution $X_1(t)$ traffic of packets $Z_1(t)$,

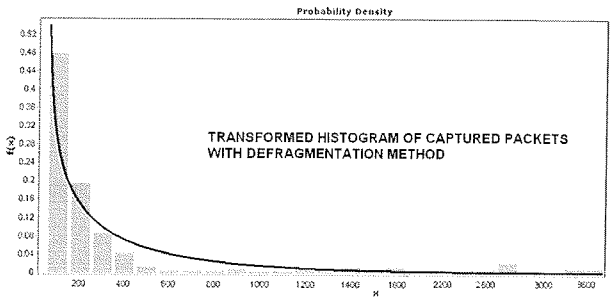

Fig. 3: Transformed histogram of captured histogram on Figure 1 with chosen distribution.

which are shorter than the maximal packets. With the second distribution $X_2(t)$, we can describe the traffic of the fragmentation packets $Z_2(t)$, which was equal to the maximal values before fragmentation.

$$Z(t) = Z_1(t) + Z_2(t) = \begin{cases} Z_1(t) = \psi_1(X_1(t), Y_1(t)); & packet\_size \neq threshold \\ Z_1(t) = \psi_2(X_2(t), Y_2(t)); & defragmentaion\_packets \end{cases} \quad (9)$$

For both processes $X_1(t)$ and $X_2(t)$ we also define and estimate distributions and their parameters for the belonging processes of inter-arrival time $Y_1(t)$, $Y_2(t)$, and also Hurst parameter, which can also be used in the modeling of arrival process.

4. Simulation results

We model the captured self-similar network traffic, which is shown on Figure 4, with short-range dependence by simulations with both classical and presented methods.

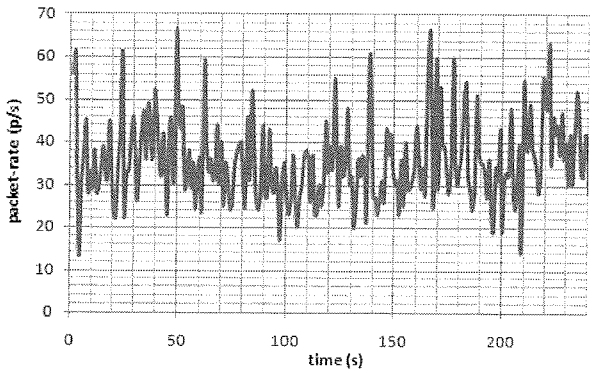

Fig. 4: Measured test network traffic captured by Wireshark sniffer.

In the case of classical estimation, we chose exponential distribution for describing the packet-size process, because the value of Hurst parameter is near 0.5 and also has short-range dependence. For the first and third methods we define threshold, which is equal to MTU size, because the bin with MTU packets is withdrawing from other neighboring local bins in the packet-size histogram (Figure 1). This bin is described by separated distribution, for the first and third methods. Table 1 shows parameters of measured network traffic and all estimated parameters for presented methods, which was used in OPNET simulations tool.

Table 1: Parameters of measured and simulated network traffic

	packet size process	inter-arrival time	p/s	kb/s	H	MSE
measured traffic	X	X	35.6	114.5	0.58	X
classical method	exponential $1/\lambda = 416{,}5$	Weibull $\alpha = 0.57326$ $\beta = 0.01895$	33.4	113.4	0.53	0.024
1. method	packets < MTU		34.1	124.9	0.54	0.016
	exponential $1/\lambda = 230.41$	Weibull $\alpha = 0.65792$ $\beta = 0.02587$				
	packets = MTU					
	constant 1482	Rayleigh $\sigma = 0.17435$				
2. method	exponential $1/\lambda = 452{,}48$	Weibull $\alpha = 0.6521$ $\beta = 0.0244$	31.0	114.5	0.52	0.026
3. method	packets < MTU		37.2	120.0	0.57	0.003
	exponential $1/\lambda = 106.7$	Weibull $\alpha = 0.677$ $\beta = 0.02932$				
	defragmentation data					
	Rayleigh $\sigma = 2181.7$	Rayleigh $\sigma = 0.1871$				

Table 1 shows the comparison between measured and modeled signals in bit and packet-rates, without method and suggested methods. There are also estimated parameters H, which are measure of a self-similarity. There are also mean square errors (MSE) between the measured and modeled histograms of the packet-size process, which also show the contents of the packets, shown in Figure 5. Using this test, we proved that presented methods impact the minimal discrepancy between measured and modeled signals and better describe measured traffic than classical estimation (without method).

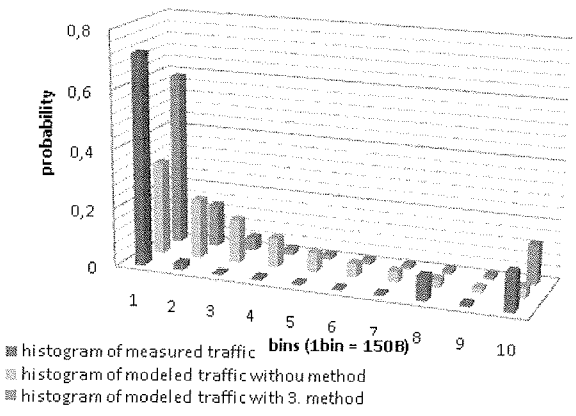

Fig. 5: Histograms of packets size process of measured traffic and modeled traffic with classical and 3.method

Figure 6 present the three simulated network traffics, which were modeled by estimated parameters from Table 1.

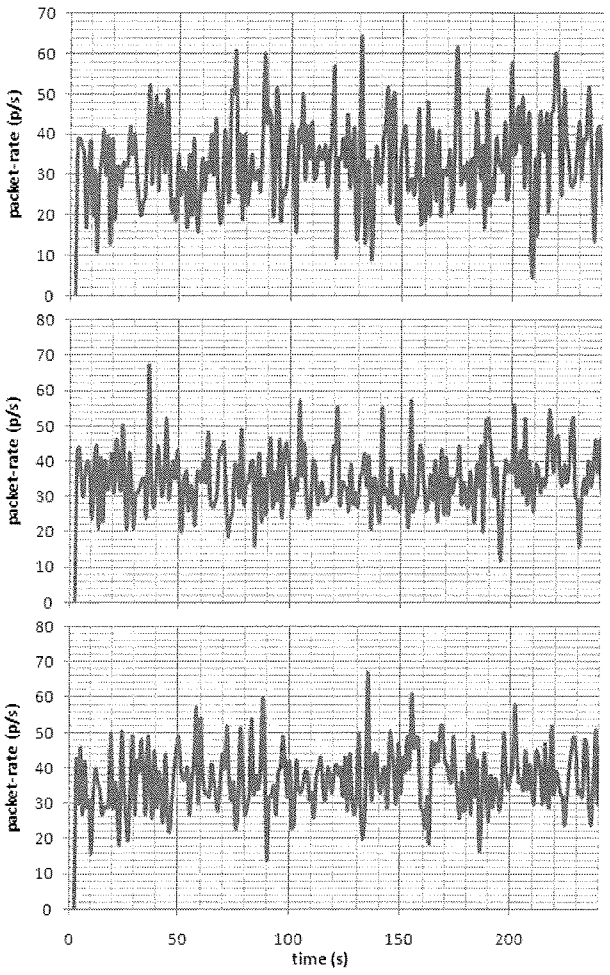

Fig. 6: Simulated network traffics in OPNET simulation tool. First graph presents simulated traffic, which was model by first method. Second graph presents simulated traffic, which was model by second method. Third graph presents simulated traffic, which was model by third method.

5. Conclusion

The presented methods show very good results in the case of modeling network traffic with short-range dependence, where we achieved better contents of packets, sometimes even better bit or packet-rates in the modeled traffic and a more accurate description of captured-traffic, then in the case of using classical manner of modeling the measured traffic. For future research we plan modeled network traffic with long-range dependence with purposeful methods, because in these cases classical estimation (without any methods) totally failed and lead to great discrepancy between measured and modeled traffic in the sense of bit and packet-rates, and also in bursts' intensities.

6. References

/1/ W. E.Leland, M. S. Taqqu, W. Willinger in D. V. Wilson, "On the self-similar nature of Ethernet traffic (Extended version)", IEEE/ACM Transactions on Networking, Vol.2, pp.1-15, 1994.

/2/ V. Paxon in S. Floyd, "Wide area traffic: the failure of Poisson modeling", IEEE/ACM Transactions on Networking, 3(3): 226-244, 1995.

/3/ K. Park in W. Willinger, "Self-Similar Network Traffic and Performance Evaluation", John Wiley & Sons, 2000.

/4/ K. Park, G. Kim in M. E. Crovella, "On the Relationship Between File Sizes Transport Protocols, and Self-Similar Network Traffic, International Conference on Network Protocols", 171-180, 1996.

/5/ W. Willinger, M. S. Taqqu, R. Sherman in D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level", IEEE/ACM Transactions on Networking, 5(1): 71-86, 1997.

/6/ F. Xue in S. J. Ben You, "The effect of aggregation on self-similar traffic", Department of Electrical and computing engineering, University of California.

/7/ J. Potemans, B. Van den Broeck, Y. Guan, J. Theunis, E. Van Lil in A. Van de Capelle, "Implementation of an Advanced Traffic Model in OPNET Modeler", OPNETWORK 2003, Washington D.C., USA, 2003.

/8/ M. Fras, J. Mohorko and Z.Cucej FRAS, "Estimating the parameters of measured self similar traffic for modeling in OPNET", IWSSIP Conference, 27.-30 June 2007, Maribor, Slovenia.

/9/ D. Persson, T. Eriksson, P. Hedelin, P," Packet Video Error Concealment With Gaussian Mixture Models", IEEE Transactions on Image Processing, Vol. 17, Issue 2, Feb. 2008 Page(s):145 - 154.

/10/ S. Luo, G. A. Marin, "Realistic internet traffic simulation through mixture modeling and a case study", Proceedings of the 37th conference on Winter simulation, 2005.

/11/ V. Karamcheti, D. Geiger, Z. Kedem and S. Muthukrishnan, "Detecting malicious network traffic using inverse distributions of packet contents", Sigcomm 2005, Philadelphia, USA.

/12/ A.Thümmler, P.r Buchholz and M. Telek, "A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM Algorithm", Dependable Systems and Networks, 2005.

/13/ H. Yölmaz, "IP over DVB: Management of self-similarity", Master of Science, Bođaziçi University, 2002.

/14/ M. Z. Jiang, "Analysis of wireless data network traffic", Master of Applied Science, Simon Fraser University, Vancouver, Canada, 2000.

/15/ O. Sheluhin, S. Smolskiy and A. Osin, "Self-Similar Processes in Telecommunications", John Wiley & Sons, 2007.

/16/ Chakravarti, Laha, and Roy, "Handbook of Methods of Applied Statistics", Volume I, John Wiley and Sons, pp. 392-394, 1967.

Matjaž Fras, Jože Mohorko, Žarko Čučej
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, 2000 Maribor, Slovenija
Epošta: matjaz.fras1@uni-mb.si

Prispelo (Arrived): 14.08.2008 *Sprejeto (Accepted): 19.03.2009*