University of Helsinki
Department of Computer Science
Series of Publications C, No. C-1994-4

Symptoms of Self-Similarity in Measured Arrival Process of Ethernet Packets to A File Server

Kimmo E. E. Raatikainen
University of Helsinki, Finland

Helsinki, April 1994

Report C-1994-4

University of Helsinki
Department of Computer Science
P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, Finland

Symptoms of Self-Similarity in Measured Arrival Process of Ethernet Packets to A File Server

Kimmo E. E. Raatikainen[†]
University of Helsinki, Finland

Recently Bellcore researchers reported that the measured Ethernet traffic has self-similar characteristics. The self-similarity calls for great changes in engineering practice when compared to traditional models for the arrival process. In this paper we report our measurement study of the arrival process of Ethernet packets into a file server. Our primary finding is that the measured index of dispersion for counts does not provide reliable evidence for or against self-similarity. This conclusion is also supported by extensive simulations.

(Traffic measurements; Arrival process; Fractal Brownian motion; Markov modulated Poisson process; Model identification)

1. Introduction

Recently Leland et al [9] reported that measured Ethernet traffic has *self-similar* characteristic or *fractal* nature. Self-similar phenomenon dates back to Hurst [8] and Mandelbrot [10]. The two main features of self-similarity are (i) structural similarities across a very wide range of time scales and (ii) hyperbolically (rather than exponentially) fast decaying autocorrelations. Detailed descriptions of mathematical and statistical aspects of self-similarity can be found for example in [1, 11]. The consequences of self-similar arrival processes are discussed in [3, 12]. The studies indicate how hopeless it is to solve traffic problems by enlarging buffers when the arrival process is self-similar.

Our primary interest is performance engineering and analysis of distributed service applications. Inspired by the results reported in [9] we wanted to learn if the arrival process of Ethernet packets to a file server has self-similar characteristics. During six weeks we measured the count of arriving Ethernet packets, daily from 9:30 to 14:30, and estimated the *index of dispersion for counts*. Our findings were quite conflicting: Some daily measurements clearly supported self-similarity while

[†]Author's address: Department of Computer Science, P. O. Box 26 (Teollisuuskatu 23), FIN-00014 University of Helsinki, Finland. E-mail:Kimmo.Raatikainen@Helsinki.FI.

some other daily measurements clearly contradicted it. About half of the daily measurements did not show obvious evidence for or against self-similar characteristics. In order to understand the phenomenon we also carried out simulation studies.

In Section 2 we briefly describe two models for the arrival process. We also discuss the characterization of traffic variability. The first model examined is *Markov Modulated Poisson Process*. The second model is *fractional Brownian motion*, which is self-similar. The measured arrival processes are described in Section 3 and simulation studies in Section 4. The concluding remarks can be found in Section 5.

2. Traffic Models

In performance engineering and analysis we need a model for the arrival process of service requests. Let $A(t)$ denote the number of arrivals in time interval $[0,t]$. In order to simplify the modeling it is usually assumed that the arrival process has stationary increments. The stationarity of increments implies that the statistical characteristics of $A(t+\delta)-A(t)$ depend solely on δ and not on t.

The index of dispersion for counts (IDC) is perhaps the best known measure for capturing the variability of traffic over different time scales. For a given time interval of length t the IDC is defined as (see e.g. [2]) the *variance-to-mean ratio* of the number of arrivals in $[0,t]$:

$$I(t) = \frac{\mathrm{Var}[A(t)]}{\mathrm{E}[A(t)]} \,. \tag{2.1}$$

The estimation of the IDC is relatively cheap from a measured arrival process, $\tilde{A}(t_i)$, $t_i = i\,\delta$, $i = 0,\ldots,n$. The primary difficulty is that we should verify that the increments, $a_i = \tilde{A}(t_i) - \tilde{A}(t_{i-1})$, $i = 1,\ldots,n$, are stationary. The estimated IDC, $\tilde{I}(t)$, for $t = k\,\delta$, is

$$\tilde{I}(k\,\delta) = \left(\overline{a_k^2} - \bar{a}_k^2\right) \Big/ \bar{a}_k \,, \tag{2.2}$$

where

$$\bar{a}_k = \frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \sum_{i=1}^{k} a_{i+(j-1)k} \qquad \text{and} \qquad \overline{a_k^2} = \frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \left(\sum_{i=1}^{k} a_{i+(j-1)k} \right)^2 .$$

The most simple model for the arrival process is Poisson, for which the increments are independent and $I(t) \equiv 1$ for every $t > 0$. However, it is reported in many measurement studies that the increments are autocorrelated. Currently the most popular model that takes the autocorrelations into account is *Markov Modulated Poisson Process* (MMPP). An MMPP is a doubly stochastic Poisson process. The arrival process is determined by the state of a continuous-time

Markov chain with K states. When the chain is in state k, the arrival process is Poisson with rate λ_k, $k = 1, \ldots, K$.

A two-state MMPP is characterized by four parameters: r_1, r_2, λ_1, and λ_2. The mean sojourn times of the chain in states 1 and 2 are $1/r_1$ and $1/r_2$, respectively. It is shown in [7] that for the two-state MMPP

$$\mathrm{E}[A(t)] = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}\, t \tag{2.3}$$

and

$$I(t) = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2(\lambda_1 r_2 + \lambda_2 r_1)} \left(1 - \frac{1 - \exp(-(r_1 + r_2)t)}{(r_1 + r_2)t}\right) . \tag{2.4}$$

A stochastic process $Z(t)$, $t \geq 0$, is said to be self-similar with the Hurst parameter H, $0 < H \leq 1$, if the increments are stationary and if $Z(\alpha t)$ is identical in distribution to $\alpha^H Z(t)$, $t \geq 0$, for every $\alpha > 0$. When the arrival process is self-similar, then the autocorrelations of increments decay hyperbolically and not exponentially as in the case of a MMPP. The hyperbolically decaying autocorrelations imply that the IDC of a self-similar process is not bounded.

The following self-similar arrival process was introduced in [12]:

$$A(t) = m\,t + \sqrt{am}\, Z_H(t)\ , \tag{2.5}$$

where m is the mean arrival rate, a is the scale factor, and $Z_H(t)$ is a *fractional Brownian motion* with the Hurst parameter H. The fractional Brownian motion $Z_H(t)$ was introduced in [11] as a moving average of the ordinary Brownian motion. The fractional Brownian motion $Z_H(t)$ has the following properties:

$$\begin{aligned} Z_H(0) &= 0\ , \\ \mathrm{E}[Z_H(t)] &= 0\ , \\ \mathrm{Var}[Z_H(t)] &= t^{2H}\ , \\ \mathrm{Cov}[Z_H(t_1), Z_H(t_2)] &= \frac{1}{2}\left(t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H}\right)\ , \end{aligned}$$

and $Z_H(t)$ is a continuous Gaussian process. Therefore, $\mathrm{E}[A(t)] = mt$ and

$$I(t) = a\, t^{2H-1} \tag{2.6}$$

for the arrival process defined in (2.5). Later we call this arrival process the *fractional Brownian motion* or fBm.

When the IDCs for the MMPP and fBm are compared, that is (2.4) and (2.6), we observe that $\log(I(t))$ for the fBm is linear in $\log(t)$ whereas $\log(I(t))$ for the MMPP is a convex function of $\log(t)$. If we plot $\log(I(t))$ against $\log(t)$, the fBm shows a linear pattern and the MMPP shows a curvature.

3. Measured Arrival Processes

As discussed in the Introduction we are primarily interested in distributed services. We measured the number of arriving Ethernet packets into one of our file servers during eight weeks. Our objective was to learn if this arrival process has self-similar characteristics. Particularly, we examined if the measured IDC supports a fBm-model or an (two-state) MMPP-model.

When we are examining the self-similar characteristics we meet the following dilemma. In order to distinguish a fBm from an MMPP each measurement period should be several hours. On the other hand, the expected traffic rate must remain (roughly) constant during the whole period, since both the fBm and the MMPP assume that the increments are stationary.

The first two weeks of the measurements were the initial phase of the study. During that time we measured the arrival rates daily from 8 am to 8 pm. Our objective was to find a period during which the expected traffic rate remains roughly constant. The measurement interval was 10 seconds. The measurement period of 12 hours was divided into non-overlapping time windows of 15 minutes. In each time window we estimated the mean arrival rate, and the 5th and 95th percentile. These two percentiles provide an approximate distribution-free 90 % confidence interval for the expected arrival rate.

Figure 1 shows an example of our initial measurements. Let m_i denote the mean, p_i the 5th and q_i the 95th percentile in the ith time window. We regard that the expected arrival rate is roughly constant from the lth time window up to the hth, if at least 90 % of $m_l, \ldots, m_h$ are in the interval $[\max(p_l, \ldots, p_h), \min(q_l, \ldots, q_h)]$. In addition, if more than two consecutive m_i's fall out of the interval, we regard that the expected arrival rate is non-constant. Based on this heuristic rule we found that usually the expected arrival rate is roughly constant from 9:30 to 14:30. In Figure 1 our rules suggest that the expected arrival rate is roughly constant from 9:15 to 15:30.

During the next six weeks (from January 31, 1994 to March 11) we estimated the IDCs from measurements, daily from 9:30 to 14:30 excluding the weekends. The measurement interval was 10 seconds. One daily measurement failed due to a system crash. Because of anomalous network traffic we rejected another measurement. Of the remaining 28 measurements we rejected three due to a "non-constant expected arrival rate".

We estimated the IDC from the daily measurements, $\tilde{I}(t_i)$ for $t_i = 10i$ and $i = 1, \ldots, 100$. A typical $\tilde{I}(t)$ shows fluctuating behaviour when the time interval t is large when compared to the measurement period T. This is not surprising since the number of non-overlapping time intervals of length t is $\lfloor T/t \rfloor$. Since our measurement period was 5 hours or 18,000 seconds, the number of time intervals on which $\tilde{I}(1000)$ is based is 18. In order to reduce the fluctuation we use a resistant technique that smooths the sequence of $\log(\tilde{I}(t_i))$'s. The smoother we use is `S4253H,twice` which is based on repeated running medians; for details see [13, Ch. 6].

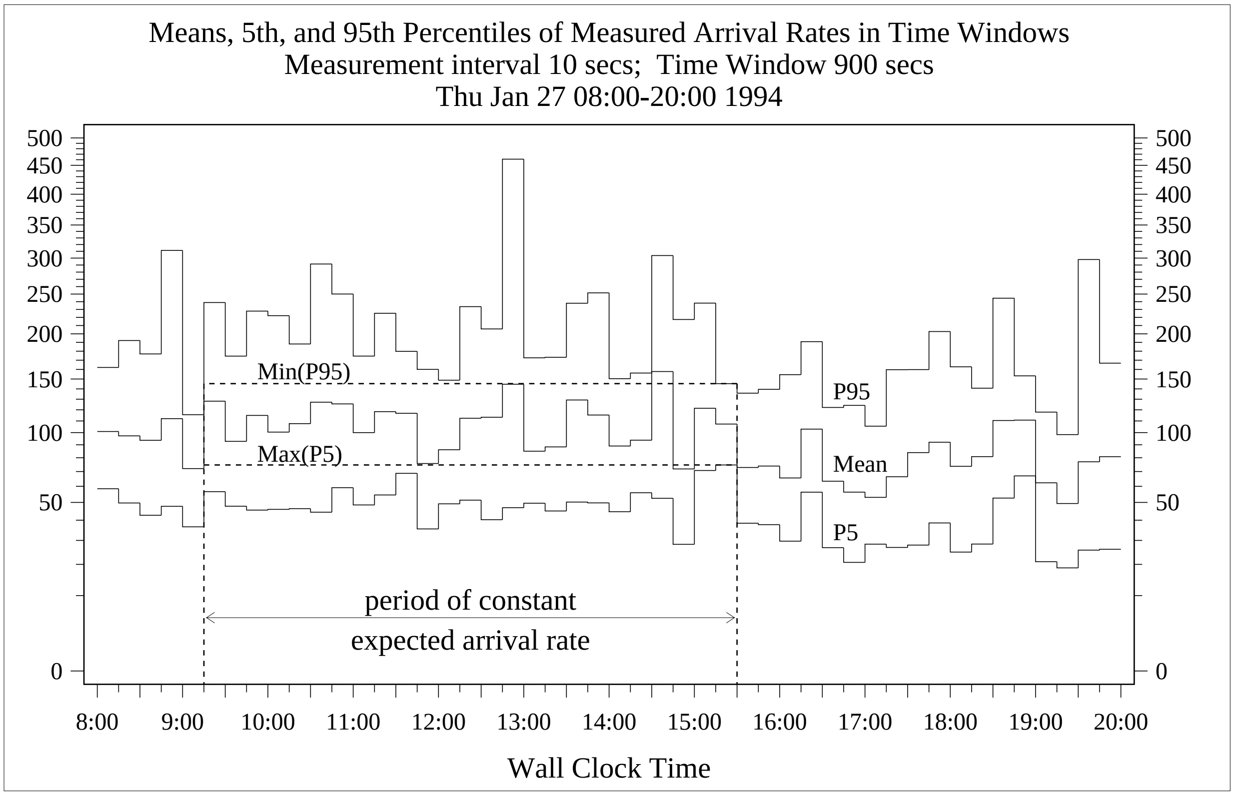


Fig. 1: Example of Daily Measurements During the Initial Phase of the Study

In order to highlight the linear or curve-like shape of $\log(\tilde{I}(t))$ as a function of $\log(t)$ we use another resistant technique. We fit a resistant regression line to the points $(\log(t_i), \log(\tilde{I}(t_i)))$ in the first $(i = 1, \ldots, 50)$, middle $(i = 26, \ldots, 75)$, and second $(i = 51, \ldots, 100)$ half of points using the algorithm given in [13, Ch. 5]. If the slopes of these three regression lines are close to each other, we conclude that the measured IDC supports self-similarity. If the slope of the regression line for the points in the second half is much smaller than for the points in the first half, we conclude that the measured IDC supports MMPP.

The 25 daily measurements show quite conflicting characteristics. Broadly speaking the measurements fell into three categories, examples of which are shown in Figure 2. Six measurements clearly supported self-similarity, that is linearity of $\log(\tilde{I}(t))$ as a function of $\log(t)$. The topmost dataset in the figure is an example of this behaviour. Seven datasets indicated an obvious MMPP-type behaviour in the way shown at the bottom of the figure. Almost half of datasets (12) indicated a behaviour that we call "don't know". The dataset in the middle is one example of this kind of pattern. Figure 3 summarizes the resistant regression lines. It shows 1) the slopes in the three resistant lines, 2) the differences between the slopes, and 3) areas (in square root scale) between the lines. This confirms our conclusion that the daily IDCs are quite different.

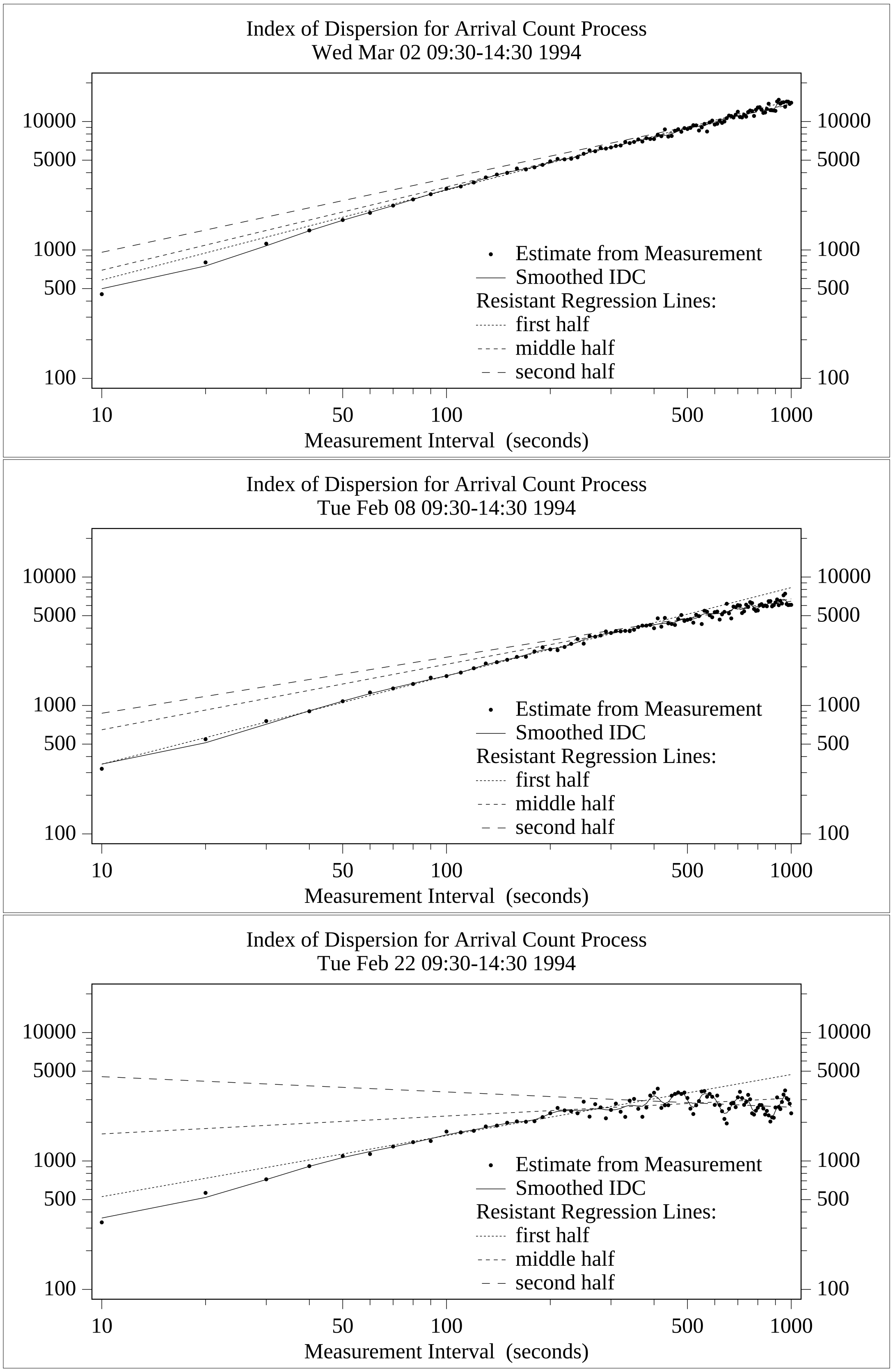


Fig. 2: Examples of IDCs Estimated from Daily Measurements

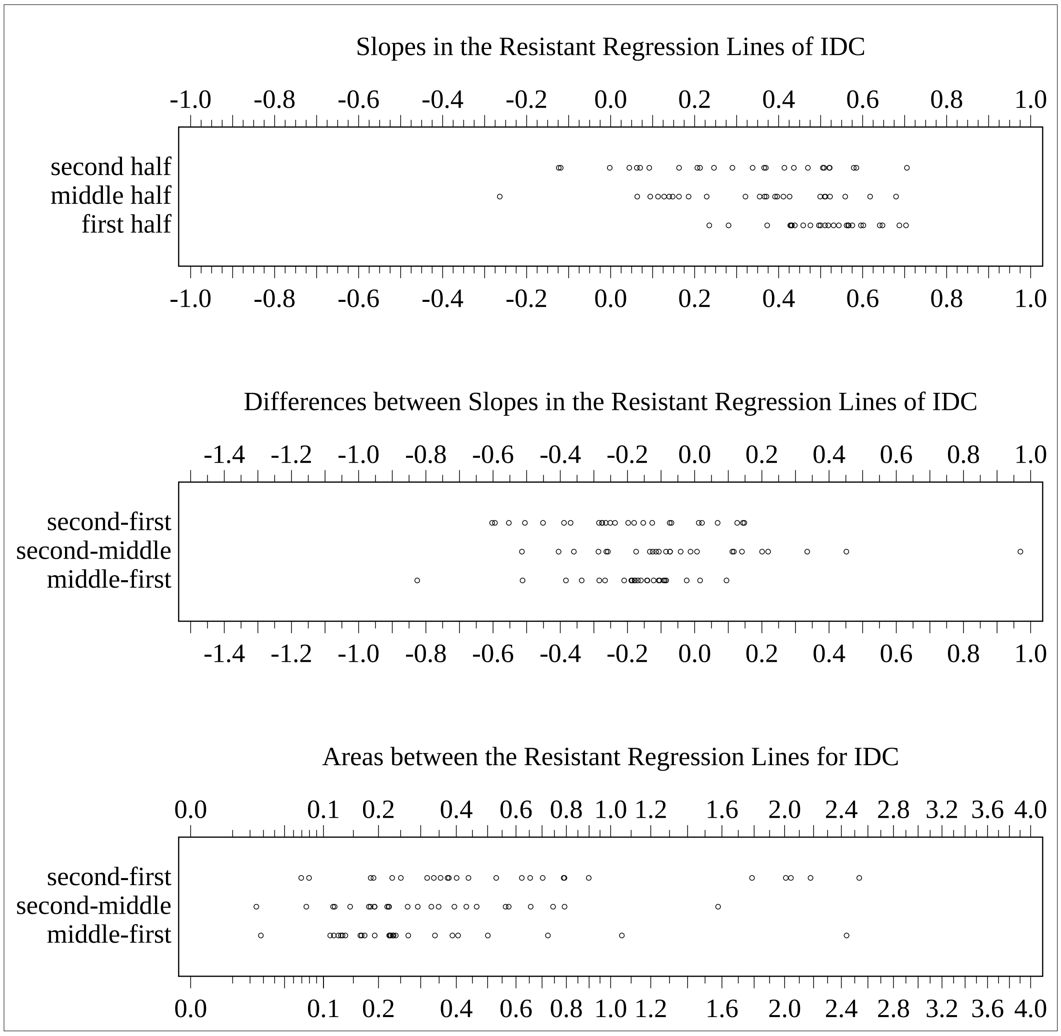


Fig. 3: Summaries Highlighting Linearity of IDCs Estimated from Daily Measurements

One way of reducing the diversity is to regard the daily measurements as independent realizations of the same stochastic process. When we take the averages of the IDCs estimated from the 25 daily measurement, we obtain an estimate of the IDC with reduced variance. Figure 4 shows the means, their smoothed values, and the three regression lines. In our classification the mean IDC falls into the category "don't known". The figure slightly, but not clearly, suggests a curved shape.

Our overall conclusion is that the measured IDCs are confusing. The measurements created several questions that we try to clarify through simulation studies. We will discuss these measurements again in Section 5 when we state our concluding remarks.

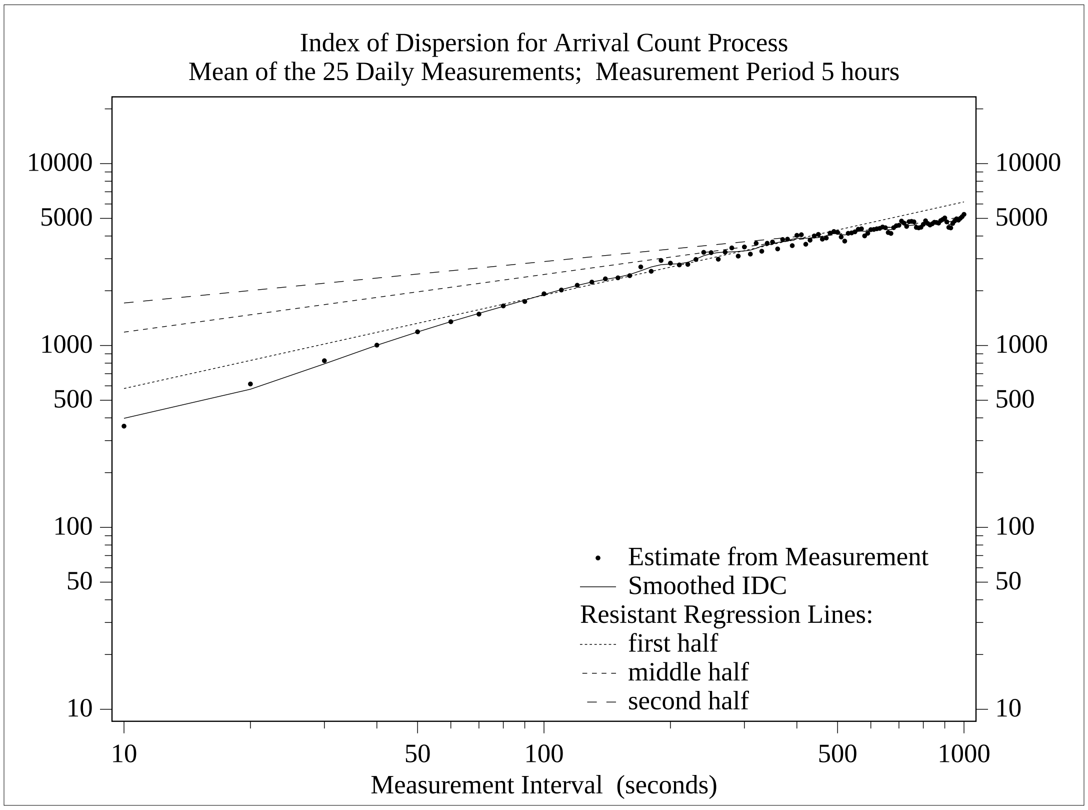


Fig. 4: IDCs Estimated from the Means of Daily Measurements

4. Simulation Studies

When we are using a stochastic model for the arrival process, we must regard a measurement as a realization of that model. Therefore we should learn the sampling properties of the indices on which the model selection and identification is based. We examined the IDCs when the arrival process is an MMPP and a fBm. Before this simulation study we examined the queue lengths at arrival instants in an A/D/1-queue, when the arrival process is an MMPP and a fBm. The objective of these arrival queue-length simulations was to learn how important it is to distinguish an MMPP from a fBm.

The arrival processes used in the simulation studies were identified in the following way. We selected one of the measurements and identified an MMPP-model and a fBm-model which minimizes

$$\sum_{i=1}^{25} \Big(I(t_i) - \tilde{I}(t_i) \Big)^2 \ ,$$

where $t_i = 10i$, $\tilde{I}(t)$ is the measured IDC and $I(t)$ is the IDC of the model given in (2.4) for the MMPP and in (2.6) for the fBm. The measured mean arrival rate, 117.812, was used as m for the

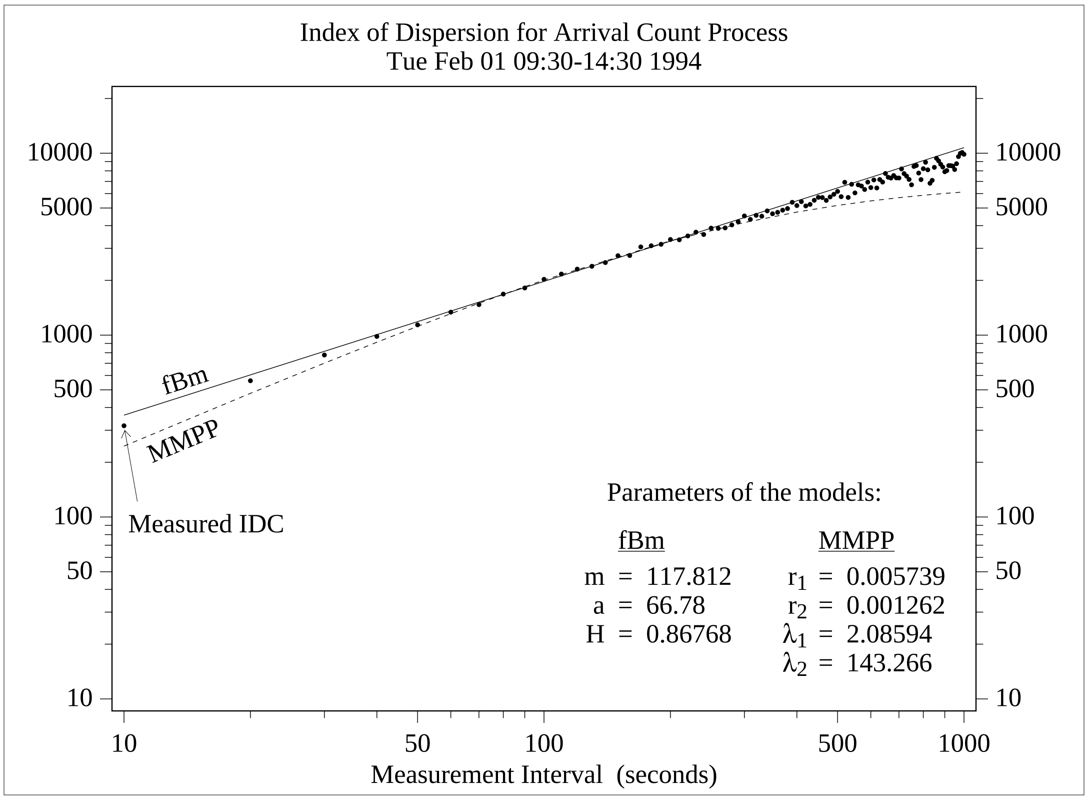


Fig. 5: Measured and Modeled IDCs

fBm. In the case of MMPP the minimization was restricted through

$$\mathrm{E}(A(1)) = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} = 117.812\ .$$

Figure 5 shows the measured IDC and the IDcs for the identified models. The parameters of the identified models are also given.

4.1 A/D/1 Simulations

In order to learn the consequences of the arrival process, we simulated an A/D/1 -queue. The deterministic service time was 4 milliseconds. Since the mean arrival rate was 117.812 packets per second, the server utilization was about 47 %. With both arrival models, MMPP and fBm, we repeated the simulation 1,000 times. The simulated time in each replication was 6 hours of which the first hour was the warmup period. During the last five hours of simulated time the number of arrivals varied from 1.7 to 2.4 millions for the MMPP and from 1.2 to 3.0 millions for the fBm. The averages were 2.116 and 2.110 millions, respectively.

In the simulations we recorded the queue length at arrival instants. Figure 6 summarizes the observed arrival queue lengths. We give the observed minimum, median, and maximum probability

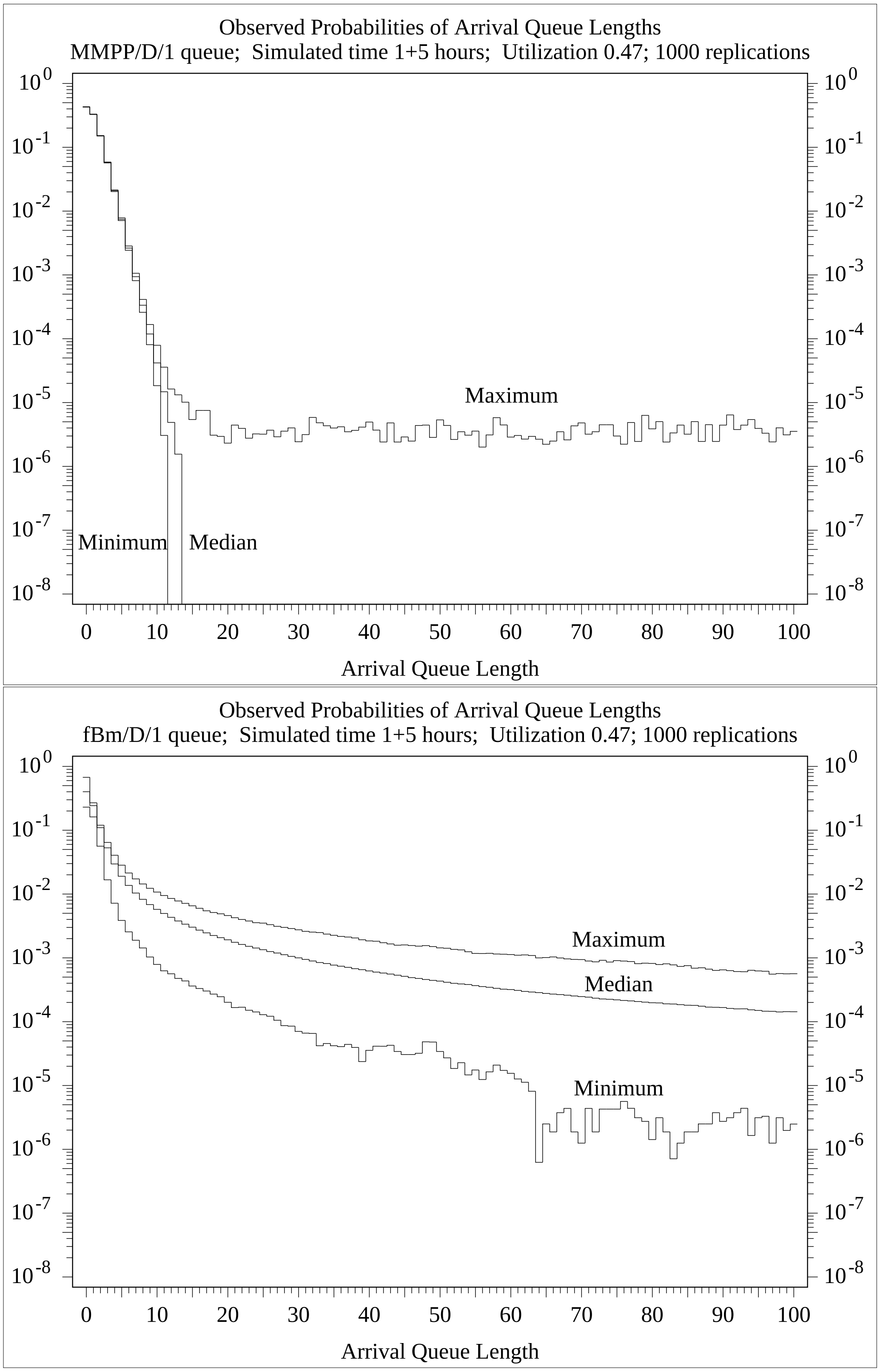


Fig. 6: Observed Probabilities of Arrival Queue Lengths

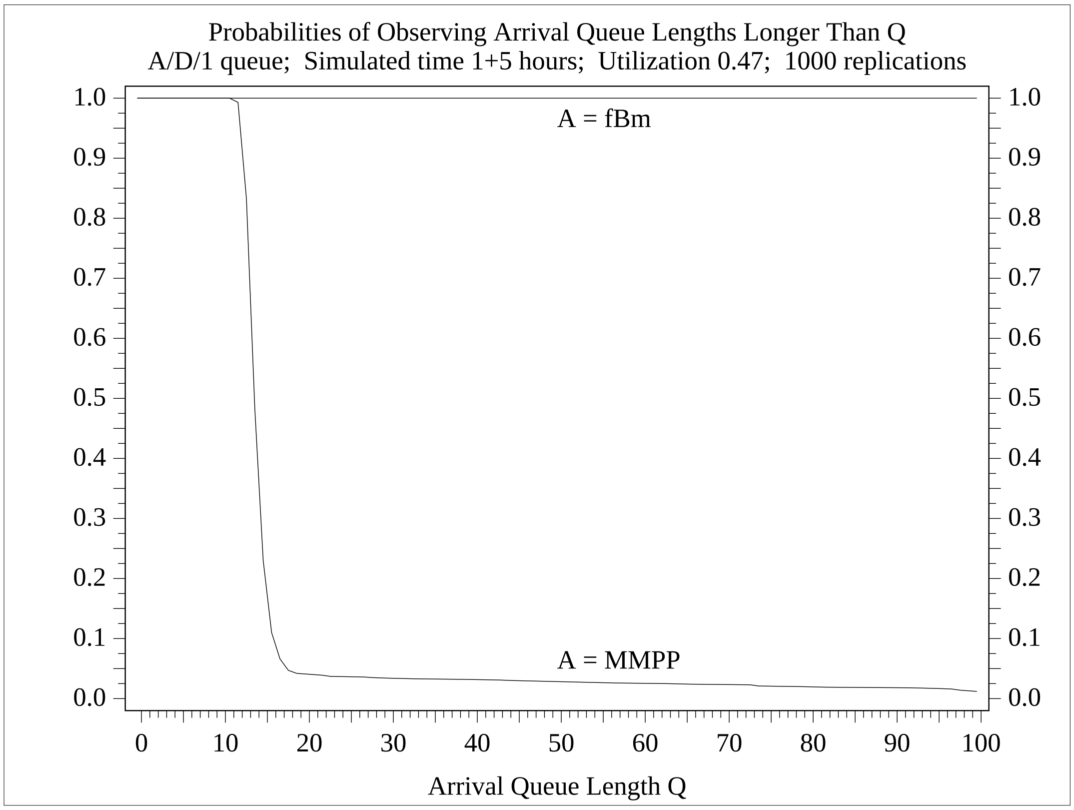


Fig. 7: Probabilities of Observing Long Arrival Queue Lengths

of the arrival queue lengths $0, \ldots, 100$. The figure clearly indicates that the queue lengths have very different characteristics. The difference has many implications. One implication is visualized in Figure 7 that gives the probabilities of observing a long queue length during five hours of simulated time.

These simulations clearly indicate that we should learn to distinguish an MMPP arrival process from a fBm. Suppose that we are using the simulations to select a suitable buffer size. If we are using an MMPP-model we must conclude that queue lengths longer than 20 are rare, since the estimated probability is $3.64 \cdot 10^{-6}$ and the estimated 99 % confidence interval is $[0, 1.74 \cdot 10^{-4}]$. On the contrary, we cannot consider queue lengths longer than 20 as rare, if our arrival model is fBm, since the estimated probability is 0.079 and the estimated 99 % confidence interval is $[0.007, 0.289]$. The confidence intervals above are based on the order statistics of the 1,000 independent replications.

4.2 Simulated Measurements

In order to learn sampling properties of the measured IDCs we generated arrival processes using the MMPP and fBm model. In addition we examined statistical techniques for distinguishing an MMPP-arrival process from a fBm. The measurement parameters are the same as in Section 3: measurement interval 10 seconds and duration 5 hours. With both models 1,000 independent replications were generated.

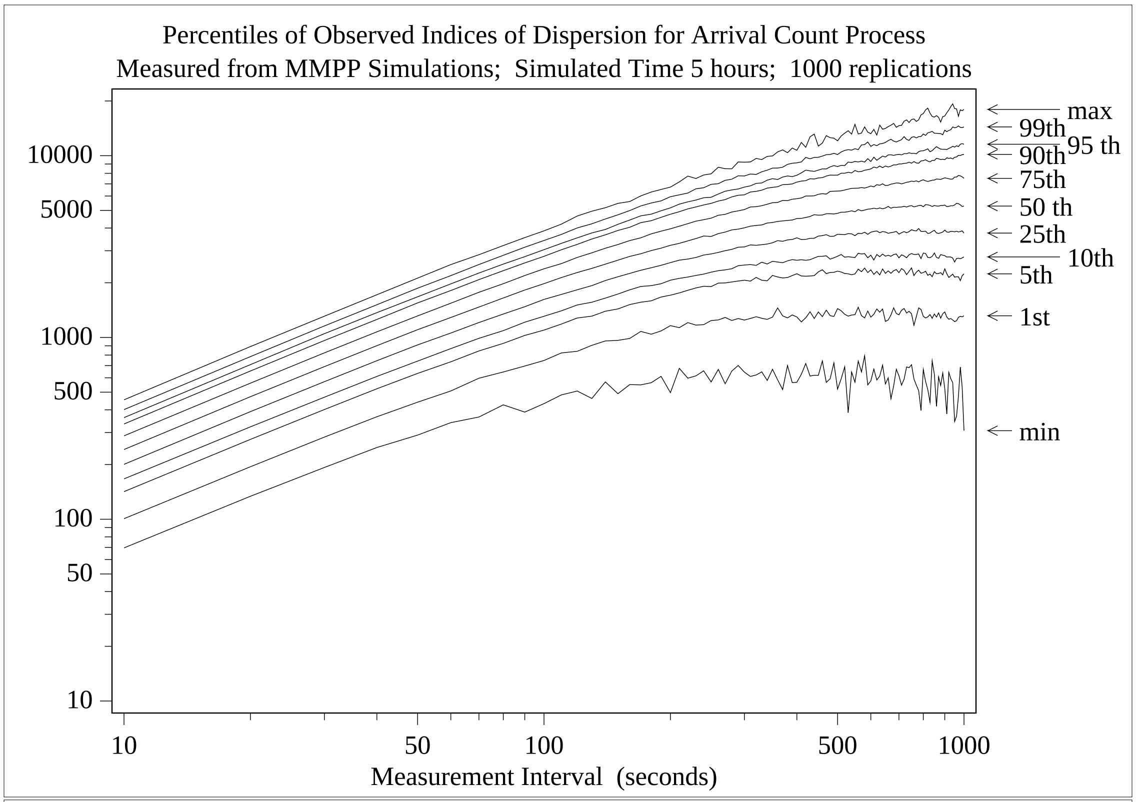
Figure 8 gives percentiles of the IDCs estimated from the simulated measurements. The percentiles clearly indicate that $\log(\tilde{I}(t))$ is linear in $\log(t)$ for the fBm and curve-shaped for the MMPP. Unfortunately, individual IDCs can show quite different shapes as Figure 9 displays. In Figure 9 the IDC from the MMPP supports (at least moderately) self-similarity and the IDC from the fBm indicate MMPP-type behaviour. Viewing through 20 randomly selected sets of 50 IDCs we found that about 40% of the IDCs clearly indicate its origin while about 20% are misleading. The rest, that is about 40%, of the IDCs do not clearly have the expected shape of an MMPP or a fBm.

In order to verify that the resistant regression lines are the summaries at which we should look when we try to distinguish an MMPP from a fBm, we examined the sampling properties of the following characteristics of the regression lines:

1. the slopes of resistant regression lines in the first, middle, and second half of the points $(\log(t_i), \log(\tilde{I}(t_i)))$: b_{f}, b_{m}, b_{s};
2. the differences between the slopes: $d_{\mathrm{m-f}} = b_{\mathrm{m}} - b_{\mathrm{f}}$, $d_{\mathrm{s-m}} = b_{\mathrm{s}} - b_{\mathrm{m}}$, and $d_{\mathrm{s-f}} = b_{\mathrm{s}} - b_{\mathrm{f}}$; and
3. the areas between the resistant regression lines: $A_{\mathrm{m-f}}$, $A_{\mathrm{s-m}}$, $A_{\mathrm{s-f}}$.

Figure 10 shows the boxplots. Notice that the boxplots of the areas are in square root scale. The square root scale was selected to improve the symmetry; see e.g. [13, Ch. 2.5]. In the boxplot bullet '$\bullet$' denotes the mean. The three quartiles are visualized by a box '$\boxed{|}$'. The whiskers '$\longrightarrow\!|$' reach from the 25$^{\text{th}}$ to the 10$^{\text{th}}$ percentile and from the 75$^{\text{th}}$ to the 90$^{\text{th}}$ percentile. The 1$^{\text{st}}$, 5$^{\text{th}}$, 95$^{\text{th}}$, and 99$^{\text{th}}$ percentile together with the smallest and largest observation are marked by ticks '|'. The boxplots indicate that we should look at the differences $d_{\mathrm{m-f}}$ and $d_{\mathrm{s-f}}$, and the square roots of areas $A_{\mathrm{m-f}}$ and $A_{\mathrm{s-m}}$, when we try to distinguish an MMPP from a fBm. It should be noted that the joint patterns of the differences and areas are not visible in the boxplots.

We also examined two statistical techniques to distinguish observations from two groups. The primary objective of these examinations was to analyze the joint effects that our summaries have when we try to distinguish an MMPP from a fBm. The first technique is the classical discrimination (or classification) described, for example, in [4, Ch. 4.1]. In this technique we must have samples from the two groups available so that we have estimates for the group means, $\hat{\boldsymbol{\mu}}_j$, and for the within-group covariance matrices, $\hat{\Sigma}_j$, $j = 1, 2$. Let n_j denote the number of observations in group j on which the estimates are based. We assign the observation $\boldsymbol{u}$ into group 1, if $\boldsymbol{a}'\boldsymbol{u} > tol$, where $\hat{\Sigma} = [(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2)/(n_1 + n_2 - 2)$, $\boldsymbol{a} = \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$, and $tol = \boldsymbol{a}'(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$. If



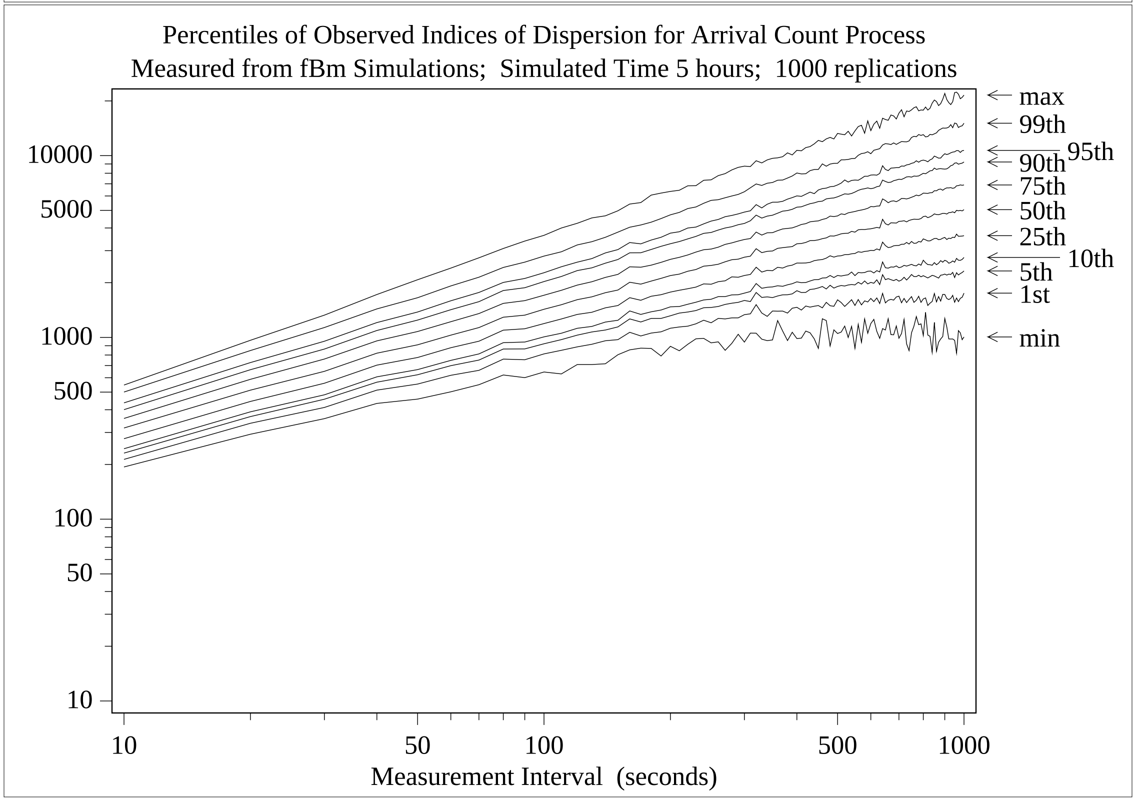


Fig. 8: Percentiles of Measured IDCs for MMPP and fBm Arrival Process

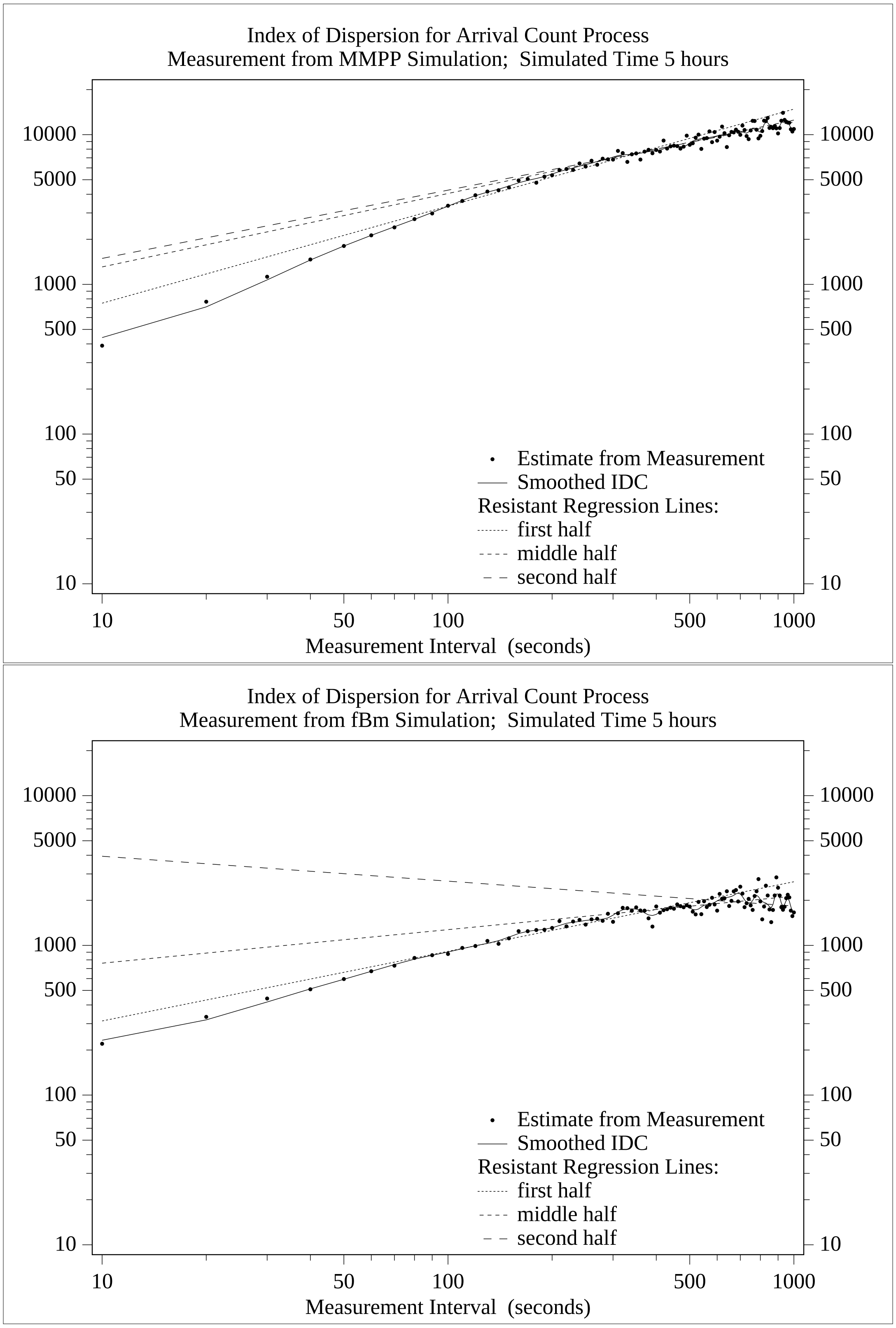


Fig. 9: Examples of IDC for MMPP and fBm Arrival Process

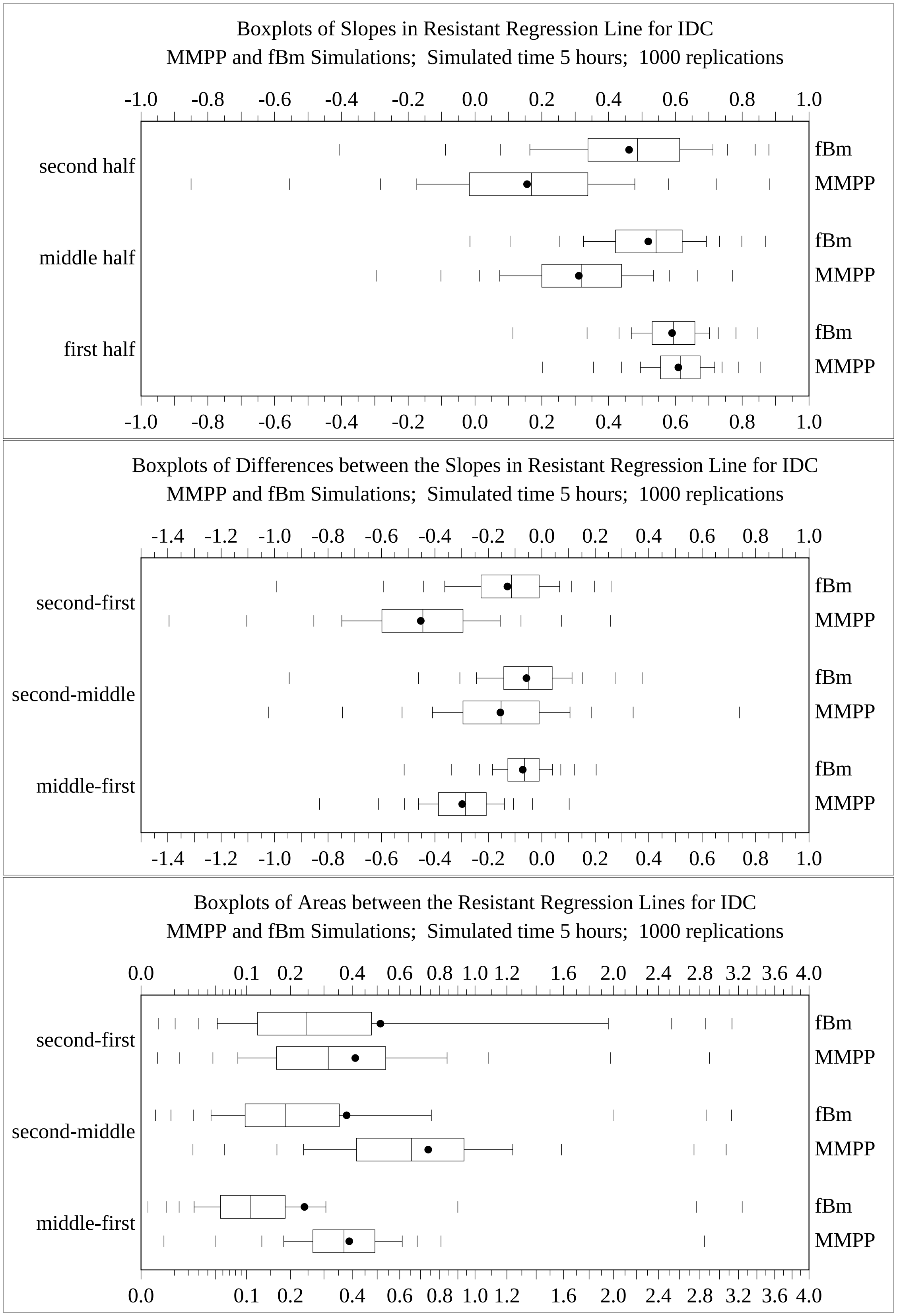


Fig. 10: Boxplots of Summaries Highlighting Difference between the IDC for MMPP and the IDC for fBm

$\boldsymbol{a}'\boldsymbol{u} < tol$, we assign the observation into group 2. In the case of $\boldsymbol{a}'\boldsymbol{u} = tol$, the group is assigned through "tossing the coin".

The second technique we use is the cluster analysis based on the *k-means algorithm* given in [6]. In this technique we assume that $k = 2$. Therefore , the only assumption we do is that the observations originates from two different groups. However, we do not have any prior knowledge about the groups.

The two techniques described above provide us two yardsticks for the probability of assigning an observation into the right group. The similarity between the true group (MMPP or fBm) and the assigned group ("MMPP" or "fBm") is one such measure. The higher the similarity is, the better discrimination is obtained by the criterion used in the classification. The similarity can be measured using the *simple matching coefficient*. Suppose that we have the following two-way classification for the number of observations:

True Group	Assigned Group "MMPP"	"fBm"	total
MMPP	m_1	$n_1 - m_1$	n_1
fBm	$n_2 - m_2$	m_2	n_2

The simple matching coefficient is $(m_1 + m_2)/(n_1 + n_2)$.

In the cluster analysis we do not necessarily obtain clusters of which one represents the MMPP and the other fBm. Therefore, we verify the results of clustering using the classical discrimination. Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ denote the centroids of the two detected clusters. If the centroids are assigned into different groups, we interpret that the clusters are representative. Otherwise, we conclude that the criterion does not provide a proper discrimination.

Table I gives the simple matching coefficients. In addition to slopes, differences, and square roots of areas, we examined the classifications based on the logarithms of the IDCs and on their smoothed values. We also examined all other combinations of (b_{f}, b_{m}, b_{s}, $d_{\mathrm{m-f}}$, $d_{\mathrm{s-m}}$, $d_{\mathrm{s-f}}$, $\sqrt{A_{\mathrm{m-f}}}$, $\sqrt{A_{\mathrm{s-m}}}$, $\sqrt{A_{\mathrm{s-f}}}$) that are not shown in the table. In classical discrimination the simple matching coefficient between the true and assigned group varied from .4635 to .787 for the combinations which have nonsingular sample covariance matrices and which are not shown in the table. In cluster analysis the coefficient varied from .517 to .7365 for the combinations which provided proper discrimination and which are not shown in the table.

The highest similarity is obtained when the classical discrimination is used and the criterion is the slopes in the first and middle halves. In cluster analysis, however, the similarity is only .707 for $(b_{\mathrm{f}}, b_{\mathrm{m}})$. In cluster analysis the highest similarity is when the criterion is the differences between the slopes in the middle and first halves and in the second and first halves. Since the similarity in

Table I: Matching Coefficients Between True and Assigned Group
The higher the coefficient is the better discrimination
between MMPP and fBm the criterion provides

Classification Criterion	Classification Method: Classical Discrimination	Classification Method: Cluster Analysis
$\log(\tilde{I}(t_i))$'s	.7145	2)
smoothed $\log(\tilde{I}(t_i))$'s	1)	2)
Slopes		
$b_{\mathrm{f}}, b_{\mathrm{m}}$	.8745	.707
$b_{\mathrm{f}}, b_{\mathrm{s}}$	.8055	.727
$b_{\mathrm{m}}, b_{\mathrm{s}}$	.761	.7625
$b_{\mathrm{f}}, b_{\mathrm{m}}, b_{\mathrm{s}}$	.5695	.75
Differences		
$d_{\mathrm{m-f}}, d_{\mathrm{s-m}}$	.8595	.715
$d_{\mathrm{m-f}}, d_{\mathrm{s-f}}$	.8595	.824
$d_{\mathrm{s-f}}, d_{\mathrm{s-m}}$	.8595	.736
$d_{\mathrm{m-f}}, d_{\mathrm{s-f}}, d_{\mathrm{s-m}}$	.8495	.7655
Sqrts of areas		
$\sqrt{A_{\mathrm{m-f}}}, \sqrt{A_{\mathrm{s-m}}}$	.7765	.6125
$\sqrt{A_{\mathrm{m-f}}}, \sqrt{A_{\mathrm{s-f}}}$	.8095	.721
$\sqrt{A_{\mathrm{s-f}}}, \sqrt{A_{\mathrm{s-m}}}$	.703	.6155
$\sqrt{A_{\mathrm{m-f}}}, \sqrt{A_{\mathrm{s-f}}}, \sqrt{A_{\mathrm{s-m}}}$	.6985	.6365

1): Sample covariance matrices were singular
2): Both clusters indicated MMPP-type behaviour

classical discrimination for any two differences is also high, we regard that $(d_{\mathrm{m-f}}, d_{\mathrm{s-f}})$ is the best criterion (of those we have examined) to distinguish an MMPP from a fBm.

We examined the twodimensional joint distribution of the $(d_{\mathrm{m-f}}, d_{\mathrm{s-f}})$'s. Our objective is to find areas in which there are many observations from one group but only a few from the other group. Our analysis is based on the idea presented in [5]. Let $\mathcal{N}$ be the set of observations: $\mathcal{N} = \{(x_i, y_i)\colon\ i = 1, \ldots, n\}$. Let $C(\mathcal{N})$ be the convex hull of $\mathcal{N}$ and let $\mathcal{N}_1$ be the subset of points in $\mathcal{N}$ that are strictly inside $C(\mathcal{N})$. Let n_1 denote the number of points in $\mathcal{N}_1$. We define that the $100 \cdot n_1/n$ percentage contour is $C(\mathcal{N}_1)$, that is the convex hull of $\mathcal{N}_1$. Recursively we obtain $100 \cdot n_j/n$ percentage contours as $C(\mathcal{N}_j)$, $j = 2, 3, \ldots$, where n_j the number of points in $\mathcal{N}_j$ and $\mathcal{N}_j$ is the subset of points in $\mathcal{N}_{j-1}$ which are strictly inside $C(\mathcal{N}_{j-1})$.

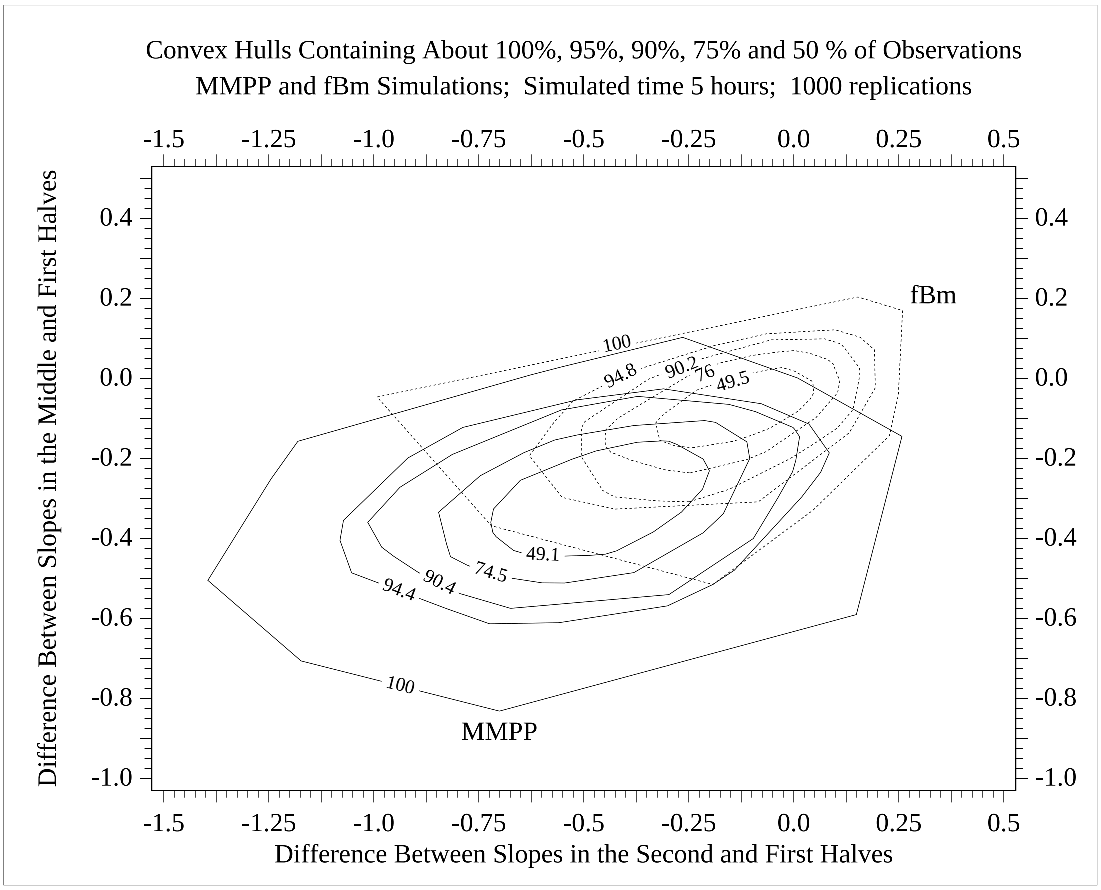


Fig. 11: Percentage Contours of Differences between Slopes
in Resistant Regression Lines of the IDC for MMPP and the IDC for fBm

Figure 11 shows some percentage contours. The figure indicates that the overlapping part of areas covered by the two groups is quite large. As the second part of this analysis we examined the cross coverages, that is the fraction of observations from one group which is inside the percentage contour of the other group. Table II gives the cross coverages for percentage contours of 100, c. 95, c. 90, c. 75 and c. 50 %. For example, 41.2 % of observations from the fBm is inside the c. 90 % percentage contour of the MMPP. The table indicate that the differences $d_{\mathrm{m-f}}$ and $d_{\mathrm{s-f}}$ for the MMPP and fBm are, at least moderately, similar.

After these examinations we repeated the viewing study with new 20 randomly selected sets of 50 IDCs. We were able to assign 76.3 % of the IDCs into the correct group. However, the success varied from 48 % to 96 % in the selected sets. Our overall conclusion from the simulated measurements is that the IDC does not provide very reliable information when we want to distinguish an MMPP from a fBm.

Table II: Cross Coverages of Percentage Contours
Fraction of observations in the other group which are inside the percentage contour of the reference group

Reference	Percentage Contour				
Group	100 %	c. 95 %	c. 90 %	c. 75 %	c. 50 %
MMPP	.82	.475	.412	.211	.075
fBm	.776	.48	.346	.173	.062

5. Concluding Remarks

We measured the arrival process of Ethernet packets into a file server during six weeks. We tried to find evidence for or against the self-similar characteristics reported in [9]. We examined the *index of dispersion for counts* estimated from measured arrival processes. Our primary finding is that the measured IDC can be confusing. Some of our daily measurements clearly supported self-similarity while some other evidently contradicted it.

Let us briefly re-examine the measurements. We present some identification results for the measured IDCs based on the simulated measurements. Figure 12 show the differences between the slopes of the resistant regression lines in the middle and first halves and in the second and first halves for the 25 daily measurements and their mean.. The percentage contours for the simulated MMPP and fBm are also shown. Table III gives the number of daily measurements assigned into the group of MMPP or fBm using classical discrimination. Table IV indicates the group into which the mean of the daily measurements was assigned. These results are yet another indication that the measured IDCs can be confusing if we try to select between an MMPP and a fBm as our model for the arrival process.

The extensive simulations indicated that the estimated IDC, $\tilde{I}(t)$, is not reliable enough to distinguish an MMPP arrival process from a fBm. The primary difficulty is that $\tilde{I}(t)$ is based only on $\lfloor T/t \rfloor$ nonoverlapping intervals. In simulation we can make the measurement period T arbitrarily long. In real measurements this is not possible since the expected arrival rate must remain roughly constant during the whole measurement period.

In [9] the degree of self-similarity was primarily estimated from the so called R/S-statistics:

$$R(n)/S(n) = \frac{\max(0, W_1, \ldots, W_n) - \min(0, W_1, \ldots, W_n)}{S(n)} ,$$

where X_k, $k = 1, \ldots, n$, is the sequence of observations, $\bar{X}(n)$ is the sample mean, $S^2(n)$ is the sample variance, and $W_k = (X_1 + \ldots + X_k) - k\bar{X}(n)$. Our primary reason of using IDC, instead of the R/S, was that the IDC is essentially easier to estimate than the R/S. We believe that

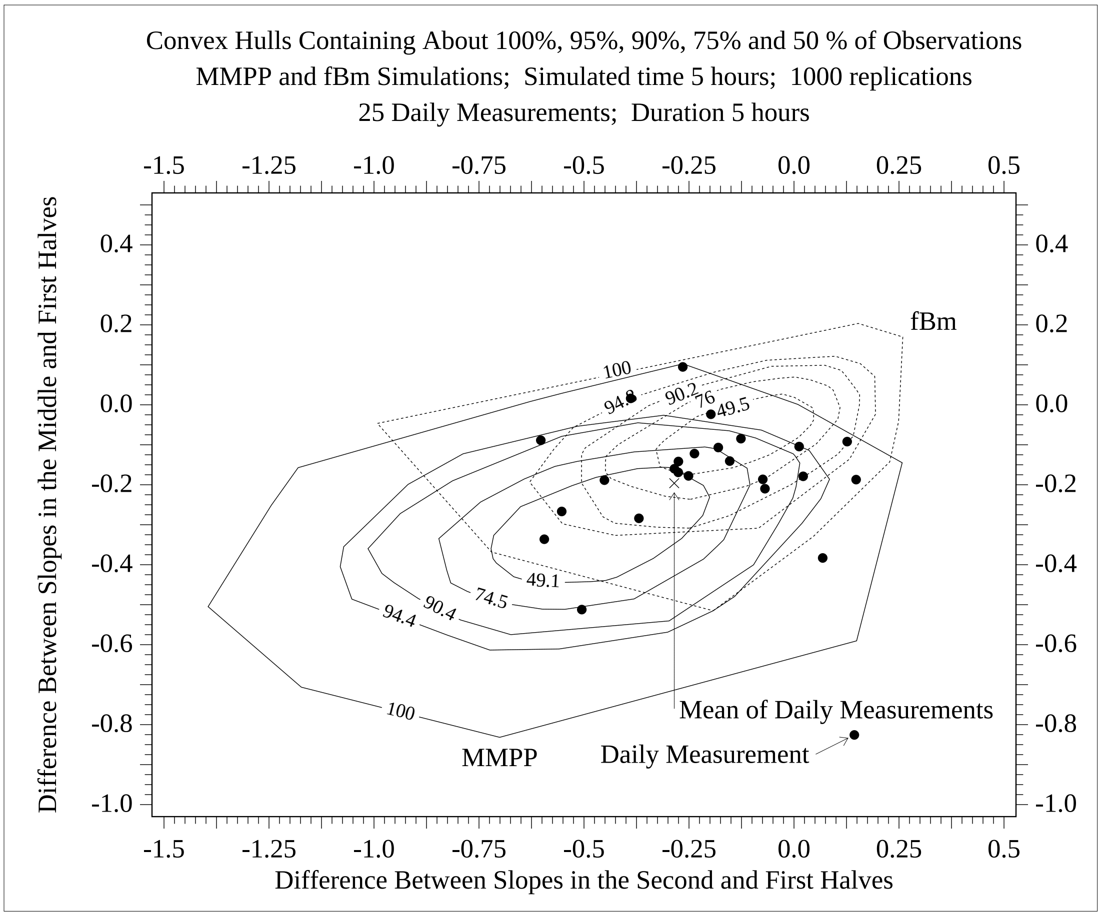


Fig. 12: Differences Between the Slopes of Resistant Regression Lines

low measurement overhead without any special purpose hardware (and kernel-level software) is essential for practical performance engineering. Our experience, however, clearly indicate that the estimation of IDC based on nonoverlapping time segments is not reliable enough.

It is proposed that overlapping time segments might be used in estimating the IDC. The advantages of these estimates are discussed for a Poisson process in [2, p. 118]. However, it is difficult to ascertain theoretically whether these advantages hold for non-Poisson processes. Nevertheless, the estimation of IDC based on overlapped time segments is simple, much easier than the estimation of the R/S-statistics. We hope that we can, in the near future, report results about estimating the IDC using overlapped time segments.

References

[1] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, Ames, Iowa, 1984.

Table III: Classification of Daily Measurements
Number of daily measurements assigned into the group of MMPP or fBm using classical discrimination and resistant regression lines

Assigned Group	Classification Criterion			
	Sqrts of areas between slopes		Differences between slopes	
	Any two	All three	Any two	All three
MMPP	11	11	7	14
fBM	14	14	18	11

Table IV: Classification of the Mean of 25 Daily Measurements

Assigned Group	Classification Criterion			
	Sqrts of areas between slopes		Differences between slopes	
	Any two	All three	Any two	All three
MMPP			X	
fBM	X	X		X

[2] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events.* Methuen & Co Lts., London, UK, 1966.

[3] H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1149, 1991.

[4] R Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations.* John Wiley & Sons, New York, 1977.

[5] P. J. Green. Peeling bivariate data. In V. Barnett, editor, *Interpreting Multivariate Data*, chapter 1, pages 3–19. John Woley & Sons, Chichester, UK, 1981.

[6] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[7] H. Heffes and D. M. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–867, 1986.

[8] H. E. Hurst. Methods of using long-term storage in reservoirs. In *Proc. of the Institution of Civil Engineers*, volume Part I, pages 519–577, 1955.

[9] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. In *Proceedings of SIGCOMM'93*, pages 183–193, San Francisco, Calif., 1993. ACM SIGCOMM.

[10] B. B. Mandelbrot. Self-similar error clusters in communication systems and the concept of conditional stationarity. *IEEE Trans. on Communications Technology*, COM-13:71–90, 1965.

[11] B. B. Mandelbrot and J. W. Van Ness. Fractional brownian motions, fractional noises and applications. *Siam Review*, 10:422–437, 1968.

[12] I. Norros. Studies on a model for connectionless traffic based on fractional brownian motion. In *Conference on Applied Probability in Engineering, Computer and Communication Sciences*, Paris, France, June 16–18, 1993. INRIA/ORSA/TIMS/SMAI.

[13] P. F. Velleman and D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, Mass., 1981.