ANALYSIS OF THE TIME EVOLUTION OF QUANTILES IN SIMULATION

MIRKO EICKHOFF, DON MCNICKLE, KRZYSZTOF PAWLIKOWSKI

University of Canterbury
Christchurch, New Zealand
E-Mail: m.eickhoff@cosc.canterbury.ac.nz

Abstract: Stochastic simulation has become a well established paradigm used in performance evaluation of various complex dynamic systems. Most simulation output analysis is confined to the estimation of mean values. This is true for both finite horizon and steady state simulation. The estimation of quantiles provides a deeper insight into the simulated model. In this paper we describe a method for estimating time evolution of several quantiles within some time interval. It is based on independent replications and its capability is demonstrated by simulating processes with different kinds of stationary, non-stationary or transient behaviour.

Keywords: Time Evolution of Quantiles, Selection of Several Quantiles, Multiple Independent Replications

INTRODUCTION

Stochastic simulation has become a well established paradigm used in performance evaluation of various complex systems, such as the Internet. Most simulation output analysis is confined to the estimation of mean values. This is true for both finite horizon and steady state simulation. Using estimators of mean values, the results of the simulation can answer questions about the average system state, such as: How many customers are there on average in the queue? The estimation of quantiles provides a deeper insight into the simulated model. Quantile estimation can additionally answer questions like: What is the probability of more than k customers in the queue? Questions of this kind are often of more interest to the decision-maker. The complexity of quantile estimation is higher than the complexity of mean value estimation, but the estimation of quantiles can give a deeper insight into the system of interest. This is true especially if several quantiles can be estimated. A set of several quantiles can be used to approximate a probability distribution function.

The most important property of a quantile estimate is its statistical accuracy. The variance of a quantile estimator usually decreases as the number of observations increases. Random errors are caused by the stochastic variations of the simulation. They are caused by the fact that every simulation is like a statistical experiment. The next source of error is the bias of the estimator itself, being often called the systematic error. This kind of error usually appears if assumptions about the analysed data hold only approximately or asymptotically. If both the variance and the bias tend to zero for large number of observations, the estimator is called consistent. More details about these statistical properties of quantile estimators can be found in [Jain and Chlamtac, 1985]. There are further properties besides these statistical ones which characterise a suitable estimator. Storage requirements and calculation time are quite important because usually a huge amount of output data needs to be processed to obtain trustworthy results. Therefore, not only the mathematical definition of the estimator, but also the way it is computed is of interest. Efficient data structures and algorithms are important. To guarantee a proper use of the estimator, even by inexperienced users, it is important that the quantile estimator is easy to understand and that the number of user-specified parameters is small, preferably zero. A classification of these properties are given e.g. in [Goldsman and Schmeiser, 1997] for the general problem of estimating standard error.

Single Quantile

The estimation of one quantile of a steady state distribution, when simulating a single instance (or "single replication") of a time-stationary process, is considered by Iglehart, Seila, Heidelberger and Lewis, Jain and Chlamtac, Chen and Kelton (see e.g. [Igelhart, 1976], [Seila, 1982], [Heidelberger and Lewis, 1984], [Jain and Chlamtac, 1985] or the more recent article [Chen and Kelton, 1999]). The methods of Igelhart and Seila are limited to regenerative processes. The subdivision of the output data into its regenerative cycles is a natural way to overcome the problem of autocorrelation. The method of Seila extends the method of Igelhart by grouping the regenerative cycles into batches. The number of parameters which have to be specified by the user is reduced by this batching approach to one parameter: the batch size. However, the determination of the

batch size is a difficulty common to every batching approach; it is difficult for an inexperienced user to choose an appropriate value. The method of Heidelberger and Welch addresses the problem of quantile estimation in dependent sequences. Their method is not limited to regenerative processes. The point estimate based on ordered data is still valid in the dependent case, but its variance is inflated leading to a larger interval estimate. Two basic solutions are given. On the one hand, the higher variance can be calculated directly with the spectral method (see [Heidelberger and Welch, 1981]). On the other hand, the data can be transformed to almost independent data by using a batch means method (see e.g. [Fishman and Yarberry, 1997]). The method of Jain and Chlamtac uses a completely different kind of quantile estimator. Their estimator is based on markers, which are adjusted when collecting new observations. This is done by a piecewise-parabolic interpolation. Because of this interpolation, this method is not recommended for quantile estimation of discontinuous distribution functions. The estimator seems to be quite complicated compared to the usual estimators based on ordered data. However, the principal advantage is that the method requires only a constant (and small) amount of memory. Chen and Kelton describe a method that estimates a quantile by focusing on observations which are located in the neighbourhood of this quantile. Their method is sequential to ensure an accurate final estimate. However, the quality of this method has not been exhaustively studied yet.

A method for quantile estimation in finite-horizon simulation is described in [Avramidis and Wilson, 1995] and [Avramidis and Wilson, 1998]. This method is based on multiple replications of the finite-horizon simulation. These replications are dependent on each other because negative correlation is introduced into their streams of input random numbers to reduced variance. Avramidis and Wilson propose that this approach yields improvements under special assumptions (see also [Jin et al., 2003]).

The estimation of one single quantile is usually done to analyse the tail behaviour of a distribution. In this case typically the 0.95-quantile (resp. 0.05-quantile) is estimated. For more extreme quantiles than this it might be more appropriate to use rare event simulation. However, sometimes the median (0.5-quantile) is estimated instead of the mean value, because the median is more robust against outliers.

Several Quantiles

If the analyst is interested in the complete distribution function of a performance measure the estimation of several quantiles is useful, because the quantiles describe the probability distribution at special points. The estimation of several quantiles of the steady state distribution is addressed by Raatikainen (see [Raatikainen, 1987]. The method of Jain and Chlamtac is extended by introducing additional markers to estimate more quantiles. The adjustment of the markers is done in the same way as before. An investigation of the variance of this method is given in [Raatikainen, 1990].

One of the main difficulties in quantile estimation is the high computational effort and the large amount of storage needed to order the observations. Therefore, Heidelberger and Welch reduce the sample size by a maximum transformation (see [Heidelberger and Lewis, 1984]). Jain, Chlamtac and Raatikainen go further and avoid sorting the output data by using an interpolation. In recent publications of Hashem, Schmeiser and Wood (see [Hashem and Schmeiser, 1994] and [Wood and Schmeiser, 1994]) or Chen and Kelton (see [Chen and Kelton, 2001] and [Chen, 2002]) quantile estimators based on order statistics have become popular again. This may be due to increased memory and processor speeds making these methods more practical. Wood and Schmeiser describe a batching method for quantiles which is similar to batch means and consider different quantile estimators, all based on ordered observations. The batch statistic is given by one of four quantile estimators, which are all based on ordered observations. Again, the difficulty is how to chose an appropriate batch size. In [Chen and Kelton, 2001] the previous method of estimating a single quantile is extended to the problem of estimating several quantiles. Again, the extended method is sequential as the previous version.

Several Quantiles Over Time

An extension of the problem of estimating several quantiles at a given time interval, or equivalent, for time-stationary processes, is analysis of the time evolution of these quantiles as the simulation progresses. This provides deeper insight into the transient behaviour of the system of interest. In steady-state simulation this can help to verify if a steady-state phase exists, i.e. that the probability distribution function of the analyzed performance measure converges to its steady-state form. The onset of the steady-state could be determined for example by the method presented in [Bause and Eickhoff, 2003].

In applications, finite-horizon simulation is frequently used to examine a given process with a certain initial state. In this case the transient behavior of the system is the central point of analysis. Again, the estimation of several quantiles over time provides a deeper insight than mean value analysis only. The application areas of quantile estimation are as vast as the application areas of simulation itself. Inventory systems, queueing systems, com-

puter systems, real-time control applications, financial industry, Internet and many more are explicitly mentioned in literature as areas of applications (see e.g. [Igelhart, 1976], [Jain and Chlamtac, 1985], [Fischer et al., 2001] or [Jin et al., 2003]).

The estimation of several quantiles in possibly time non-stationary processes has had limited attention. In the following section the use of multiple replications for this topic is discussed. Then two alternative methods are discussed, which are able to select a suitable number of quantiles based on their confidence intervals. The better method is used to examine examples with a variety of different transient behaviors. Conclusions are given in the last section.

INDEPENDENT REPLICATIONS

The main problem in quantile estimation is that the output data $X_1, X_2, \ldots$ of a single simulation run is typically non-stationary and autocorrelated (see e.g. [Lee et al., 1999]). Therefore, the amount of output data required can be immense, which causes a problem when storing and sorting the output data. Using p independent replications of the simulation is a well known approach to obtain independent sequences of output data. Let $\left\{\{x_{j,i}\}_{i=1}^{n_j}\right\}_{j=1}^{p}$ denote the collected observations. $x_{j,i}$ is the ith observation of the jth replication. n_j is an unbounded value which denotes how many observations are collected in the jth replication. With limited loss of generality we will assume that $\forall j : n_j = n$. Additionally, let us assume that $\forall j : X_{j,i} = X_i$ holds for a constant value of i, where $X_{j,i}$ is the random variable of the observation $x_{j,i}$. This means that the ith observations of all replications describe the same (possibly) transient measure. For example the ith observation could be the delay of the ith customer leaving a system, or it could be defined as the queue length at model time $i \cdot 100$ seconds. These assumptions ensure that the data in the ith column is independent and identically distributed, i.e.

$$\Pr\left[\forall j : X_{j,i} \leq x\right] = \prod_j \Pr\left[X_{j,i} \leq x\right]$$

and

$$\forall j : F_{X_{j,i}}(x) = F_{X_i}(x),$$

respectively. $F_{X_i}(x) = \Pr\{X_i \leq x\}$ denotes the cumulative probability distribution function (CDF) of a random variable X_i. In [Bause and Eickhoff, 2003] these assumptions are used to determine a truncation point for steady state simulation.

Here, these assumptions allow us to estimate the cumulative probability distribution $F_{X_i}(x)$, by

$$\hat{F}_{X_i}(x) = \frac{1}{p}\sum_{j=1}^{p} \zeta(x - x_{j,i}) \qquad (1)$$

with

$$\zeta(\Delta) = \begin{cases} 1, & \text{if} \quad \Delta \geq 0, \\ 0, & \text{else.} \end{cases}$$

$\hat{F}_{X_i}(x)$ is called the empirical cumulative distribution function (ECDF). The value of $\hat{F}_{X_i}(x)$ is determined by counting how many observations of $\{x_{j,i}\}_{j=1}^{p}$ are smaller than x. If k values of $\hat{F}_{X_i}$ are of interest, the use of Equation (1) leads to a time complexity of $\mathrm{O}\,(kp)$, which is quite inefficient. In this situation it is advisable to base the estimation on a sorted random sample. Let $\{y_{j,i}\}_{j=1}^{p}$ be the ordered sequence of $\{x_{j,i}\}_{j=1}^{p}$. Equation (1) can be changed to

$$\hat{F}_{X_i}(x) = \frac{1}{p}\min(j | x \geq y_{j,i}) \qquad (2)$$

with $1 \leq j \leq p$ and $\hat{F}_{X_i}(x) = 0$ for $x < y_{1,i}$. The calculation of k points of $\hat{F}_{X_i}(x)$ from Equation (2) can be done in $\mathrm{O}\,(k + p\log p)$, because the data has to be sorted only once. Furthermore, each value $y_{j,i}$ is an estimate of the q-quantile of F_{X_i} at $q = j/p$.

The q-quantile of the cumulative probability distribution F_{X_i} is defined by $q = F_{X_i}(x_q)$ and therefore,

$$x_q = F_{X_i}^{-1}(q) = \inf\{x | F_{X_i}(x) \geq q\}$$

is the location of the q-quantile in the case of a continuous distribution $F_{X_i}(x)$. A valid estimator for the location of the q-quantile at observation index i is given by

$$\hat{x}_q = y_{\lceil pq \rceil, i}. \qquad (3)$$

To simplify the notation, the dependence on i is omitted on the left side of the equation. The half width of a confidence interval of $\hat{x}_q$ can be described in two ways:

$$\text{either as} \quad \hat{x}_q \in x_q \pm \epsilon'_q \quad \text{or} \quad \hat{x}_q \in x_{q \pm \epsilon_q}.$$

ϵ'_q describes an interval in the range of the measure and ϵ_q describes an interval in the range of the probability (see [Chen and Kelton, 1999]). Note, the interval $q \pm \epsilon_q$ should not exceed the bounds 0 and 1. ϵ'_q and ϵ_q are dependent on each other. If one is decreased, e.g. ϵ'_q, the related ϵ_q will decrease automatically. However, in steep areas of F_{X_i} we expect ϵ'_q to be smaller (relatively) than ϵ_q. In flat areas of F_{X_i} we expect ϵ'_q to be bigger (relatively) than ϵ_q. This is demonstrated in Figure 1 with the example of an exponential distribution. Note, the difference between the steep and the less steep regions of the curve.

In general, ϵ'_q can be calculated from

$$\begin{aligned} \Pr\{y_{l,i} \leq x_q < y_{u,i}\} &= 1 - \alpha_{l,u} \qquad (4) \\ &= \sum_{j=l}^{u-1} \binom{p}{j} q^j (1-q)^{p-j} \end{aligned}$$

by decreasing l and increasing u until the chosen confidence level $(1 - \alpha) \leq$

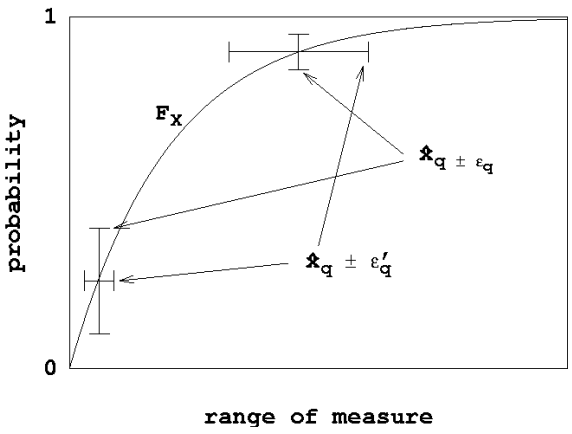


Figure 1: Confidence intervals for quantiles.

$(1 - \alpha_{l,u})$ is reached (see [Conover, 1999] and [Heidelberger and Lewis, 1984]). l and u are both ranks in the ordered sample $\{y_{j,i}\}_{j=1}^{p}$ of the original observations $\{x_{j,i}\}_{j=1}^{p}$ and describe the location of the lower and the upper border of the confidence interval. They should not exceed the rank borders of 0 and p. Note that neither the value of the lower border $y_{l,i}$ nor the value of the upper border $y_{u,i}$ are involved in the calculation of the Equation (4). Theoretically, the distributions of order statistics are not symmetric, in general. However, in our calculations we always decrease l in the same way as we increase u, therefore our obtained confidence intervals are symmetric.

In [Chen and Kelton, 1999] it is demonstrated that ϵ_q can be chosen from the inequality

$$p \geq \frac{z_{1-\alpha/2}^2 q(1-q)}{\epsilon_q^2}, \qquad (5)$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. ϵ_q can be calculated in dependence of p, q and $z_{1-\alpha/2}$. The confidence level α can be regarded as a constant parameter, and hence $z_{1-\alpha/2}$. q defines the quantile itself. p remains as the only important parameter. Note, ϵ_q does not depend on the collected observations.

Both, Equation (4) and Inequality (5) do not depend on the output data itself. Therefore, both formulas can be used to estimate the half width before the simulation experiment starts. From the point of view of mean value analysis, this is quite surprising as a confidence interval for an estimated mean value depends on the output data itself. However, Equation (4) and Inequality (5) mainly depend on the number of replications p, because the confidence level $1 - \alpha$ can be considered in both cases as a constant parameter. Therefore, p is the most important parameter in the methods described in subsequent sections.

To fully investigate the transient behaviour of a measure of interest an analysis of several quantiles over time is needed. As discussed above, the use of independent replications enables the estimation of $F_{X_i}(x)$ based on the j/p-quantiles. However, is it really appropriate to use all of these $1/p, \cdots, j/p, \cdots, 1$ quantiles to e.g. depict the transient behaviour? Because the confidence intervals of two adjacent quantiles at j/p and $(j+1)/p$ overlap extensively it is questionable to use both quantiles. To allow a clear depiction the quantiles should be chosen with non-overlapping confidence intervals. This suggests a method which determines a maximum number of quantiles with non-overlapping confidence intervals, for a given number of replications p, because the half width of the confidence interval of $\hat{x}_q$ depends on p.

SELECTION OF QUANTILES

As already noted, the calculation of the confidence interval of the q-quantile based on Equation (4) and Inequality (5) does not depend on the output data itself, but on the number of replications p, the confidence level $1 - \alpha$ and q itself. Because the confidence level can be considered as a given parameter the main question is: How to choose several q-quantiles as a function of p? The basic idea of the algorithms described in this section is to choose the 0.5-quantile as the starting point and to choose all other quantiles in a way that their confidence intervals do not overlap. A larger number of replications will produce smaller confidence intervals and this enables the selection of more quantiles with non-overlapping confidence intervals.

Rank Domain

Our first method is based on Equation (4). In the beginning the first quantile 0.5 is estimated and its confidence interval is calculated by extending l and u until the desired confidence level $1 - \alpha$ is reached. l and u describe the indexes in $\{y_{j,i}\}_{j=1}^{p}$ of the bounding values of the confidence interval. The selection of the next two quantiles which have non-overlapping and non-disjoint confidence intervals is not straight forward because Equation (4) has no closed form. Therefore, we perform two binary searches in the directions above and below 0.5. The binary search in the direction below 0.5 stops if a quantile is found with an upper bound u' being equal to l. Analogously, the binary search in the upper direction stops if a quantile is found with a lower bound l' equal to u. These binary searches give the next quantiles. The binary searches are repeated until it is not possible to find another quantile with a confidence interval in the unprocessed area between the last l and 1 (resp. u and p). This calculation can be performed before the simulation experiment starts, and therefore, the run time of this method does not really matter. For convenience a linear search, leading to a worse run time, could be

performed instead of the binary search.

An example of the binary search is depicted in Figure 2. The first selected quantile is $q = 0.5$, which is located at the rank $0.5 \cdot p$ in the ordered sequence. Its confidence interval can be calculated by Equation (4), and so the lower bound located at the rank $l_{p/2}$ is known. The unprocessed area reaches now from 0 to $l_{p/2}$. In the first step of a binary search this area is divided in two. The quantile, which is placed at the midpoint is calculated by Equation (3). Its confidence interval is then calculated by Equation (4). In this example there is a gap between the new and the current confidence interval. Therefore, the new estimated quantile is not adequate. An adequate quantile must be located in the right half between $0.5 \cdot l_{p/2}$ and $l_{p/2}$. The binary search is continued in a second step by dividing the area between $0.5 \cdot l_{p/2}$ and $l_{p/2}$ in two. In this step the new confidence interval overlaps. The binary search is continued until a quantile with a non-overlapping and non-disjoint confidence interval is found. The next binary search is started for the next quantile. This is continued until it is not possible to place another quantile in the unprocessed area, i.e. the confidence interval is too wide. Note, all values of l and u describe ranks and must be rounded, if necessary.

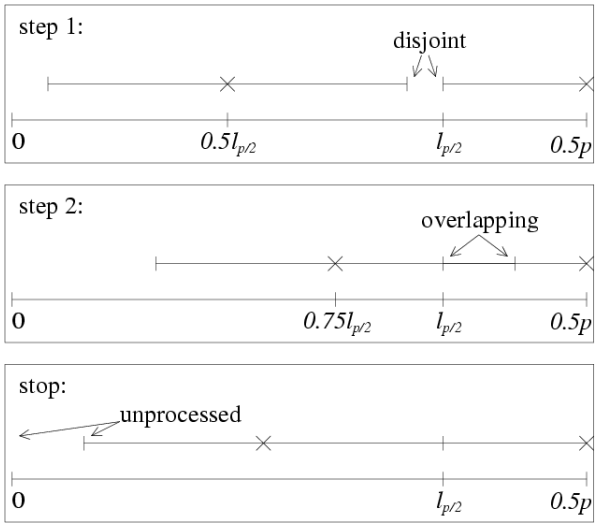


Figure 2: Binary search in the rank domain to select quantiles with non-overlapping and non-disjoint confidence intervals.

The first two columns of Table 1 show the rounded results of this method for $p = 100$ and $p = 1000$ independent replications with a confidence level of $1 - \alpha = 0.9$. The method selects seven quantiles for $p = 100$ and 27 quantiles for $p = 1000$. The values in brackets show the upper bound l and the lower bound u. These two values are ranks in the ordered sequence and define the confidence interval based on ϵ'_q (see Equation (4)).

Equ. (4)		Inequ. (5)	
$p = 100$	$p = 1000$	$p = 100$	$p = 1000$
q (l;u)	q (l;u)	q ($q \pm \epsilon_q$)	q ($q \pm \epsilon_q$)
			.003 (0;.006)
	.010 (5; 16)		.012 (.006;.018)
	.024 (16; 32)		.026 (.018;.034)
	.042 (32; 53)		.045 (.034;.056)
.08 (3;13)	.066 (53; 79)	.09 (.04;.13)	.069 (.056;.082)
	.094 (79;110)		.098 (.082;.113)
	.127 (110;145)		.131 (.113;.148)
.19 (13;26)	.164 (145;184)	.20 (.13;.26)	.167 (.148;.187)
	.205 (184;226)		.208 (.187;.230)
	.250 (227;273)		.252 (.230;.274)
.34 (26;42)	.297 (273;321)	.34 (.26;.42)	.298 (.274;.322)
	.346 (321;371)		.346 (.322;.371)
	.396 (371;422)		.397 (.371;.422)
	.448 (422;474)		.448 (.422;.474)
.5 (42;59)	.5 (474;527)	.5 (.42;.58)	.5 (.474;.526)
	.553 (527;579)		.552 (.526;.578)
	.604 (579;630)		.603 (.578;.629)
	.655 (630;680)		.653 (.629;.678)
.67 (59;75)	.704 (680;728)	.66 (.58;.74)	.702 (.678;.726)
	.751 (729;774)		.748 (.726;.771)
	.795 (774;817)		.792 (.771;.813)
.81 (75;88)	.836 (817;856)	.80 (.74;.87)	.833 (.813;.852)
	.873 (856;891)		.869 (.852;.887)
	.906 (891;922)		.902 (.887;.918)
.93 (88;97)	.935 (922;948)	.91 (.87;.96)	.931 (.918;.944)
	.958 (948;969)		.955 (.944;.966)
	.977 (969;985)		.974 (.966;.982)
	.990 (985;996)		.988 (.982;.994)
			.997 (.994; 1)

Table 1: Selected quantiles chosen by Equation (4) and Inequality (5) with $1 - \alpha = 0.9$, for $p = 100$ and $p = 1000$, respectively.

Probability Domain

The second method we investigate is based on Inequality (5). Again, the starting point is the 0.5-quantile and the method searches for more quantiles in the directions below and above 0.5. In this case a binary search is not needed, because the next quantile can be calculated directly with the help of Inequality (5) and the following conditions:

$$q_k < 0.5: \quad q_k - \epsilon_{q_k} = q_{k+1} + \epsilon_{q_{k+1}} \quad (6)$$

$$q_k > 0.5: \quad q_k + \epsilon_{q_k} = q_{k+1} - \epsilon_{q_{k+1}} \quad (7)$$

q_k denotes the kth selected quantile. The first condition is valid for the direction below the probability 0.5 and ensures that the upper bound of the confidence interval of the current quantile is equal to the lower bound of the previous confidence interval. The second condition is valid for the direction above 0.5. It ensures that the lower bound of the new confidence interval is equal to the upper bound of the previous confidence interval. In the following we focus on the first condition, because the second condition can be treated analogously. We can assume that q_k is given or already calculated, because in the beginning we choose $q_0 = 0.5$. ϵ_{q_k} can be calculated by Inequality (5). Therefore, we can use the substitution $a_k = q_k - \epsilon_{q_k}$.

Equation (6) can be transformed to:

$$a_k = q_{k+1} + z_{1-\alpha/2}\sqrt{\frac{q_{k+1}(1-q_{k+1})}{p}}$$

Eliminating the square root leads to

$$0 = q_{k+1}^2 b + q_{k+1} c_k + d_k$$

with $b = \frac{1}{z_{1-\alpha/2}^2} + \frac{1}{p}$, $c_k = -\frac{2a_k}{z_{1-\alpha/2}^2} - \frac{1}{p}$ and $d_k = \frac{a_k^2}{z_{1-\alpha/2}^2}$. Finally, the new quantile q_{k+1} can be calculated by

$$q_{k+1} = \frac{-c_k - \sqrt{c_k^2 - 4bd_k}}{2b}. \qquad (8)$$

Equation (8) is valid for quantiles below 0.5 and its use is demonstrated in Figure 3. An equation for quantiles above 0.5 can be derived analogously. Furthermore, the location of the selected quantiles is symmetric with the centre 0.5, except for errors due to rounding of non integer values. The selection of more quantiles is continued until the bounds of the probability domain $[0, 1]$ are exceeded. With this approach the probability domain is filled with non-overlapping and non-disjoint confidence intervals.

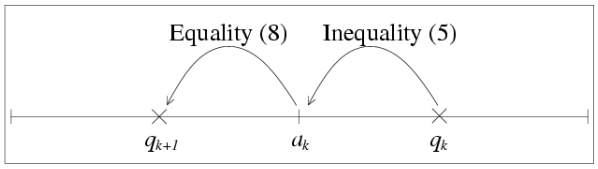


Figure 3: Selection of quantiles with non-overlapping and non-disjoint confidence intervals in the probability domain.

The rounded results of the second method are shown in the last two columns of Table 1. For $p = 100$ the second method selects seven quantiles and for $p = 1000$ this method selects 29 quantiles. The values in brackets show the confidence interval of the belonging quantile in the probability domain.

Comparison

A comparison of the results of the first and second method can be done by transforming the rank domain to the probability domain, or vice versa: i.e. by dividing the rank by p or multiplying the probability by p. The results of the first and the second method are comparably accurate. However, the binary search of the first method is complex compared to the direct calculation by Equation (8) in the second method. Furthermore, the calculation of $\binom{p}{j}$ in Equation (4) involves the handling of very small and very large values. This might lead to problems in computer calculations and rounding errors. Therefore, we recommend the second method. All the examples in subsequent sections use quantiles selected by the second method.

EXAMPLES

In the previous section we described how to select a number of quantiles. In this section we use these quantiles to investigate three stochastic processes (see Figure 4) with known statistic properties, as in [Bause and Eickhoff, 2003]. This is followed by an investigation of the time evolution of quantiles of more complex models (see Figures 5, 6, 7 and 8). These investigations show that the transient behaviour of quantiles is a very intuitive way to depict the transient behaviour of a given process.

In all our simulations we used the random number generator described in [L'Ecuyer et al., 2002]. This generator allows the choice of many substreams, making it suitable for multiple independent replications. In all experiments demonstrated in this section we used $p = 1000$ replications and the selection of several quantiles within the probability domain as described in the previous section. In [Eickhoff et al., 2005] some experiments are done with $p = \{50, 100, 500\}$ independent replications leading to a smaller number of estimated quantiles.

An ARMA Process

Autoregressive moving average (ARMA) processes are commonly used in time series analyses. They are a class of stochastic processes with well known statistical properties. To validate our method of transient quantile estimation we use an ARMA$(5, 5)$ process which is defined by

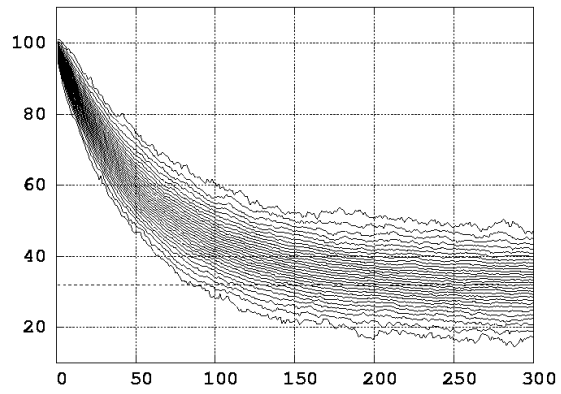
$$X_i = 1 + \epsilon_i + \sum_{k=1}^{5} \frac{1}{2^k}(X_{i-k} + \epsilon_{i-k}), k \geq 0$$

with the starting condition $X_{-5} = X_{-4} = X_{-3} = X_{-2} = X_{-1} = 100$. $\{\epsilon_i\}_{i=1}^{\infty}$ is an independent Gaussian white noise process ([Hamilton, 1994]). Therefore, the process is normally distributed for any i with a transient mean and variance. The expected value of this process for large i is $\mathrm{E}[X_\infty] = 32$ (see the dashed line in Figure 4(a)). This process is highly autocorrelated, because its current value depends on five previous values. The process is expected to converge from the initial value 100 to 32. The estimates of the transient quantiles are shown in Figure 4(a). The simulation of the ARMA process behaves exactly as expected. Additionally, we get an impression of the speed of the convergence, which is high in the beginning and is increasing with decreasing i.

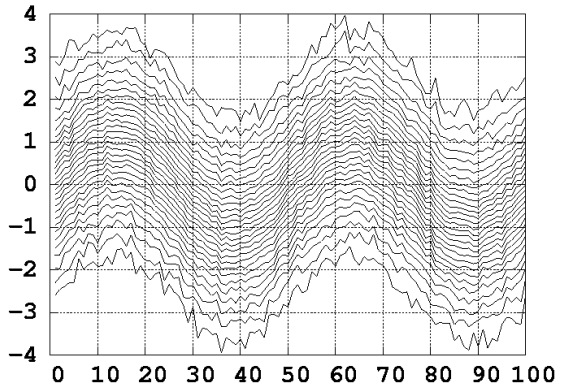
A Periodic Process

The second examined stochastic process is periodic and is defined by

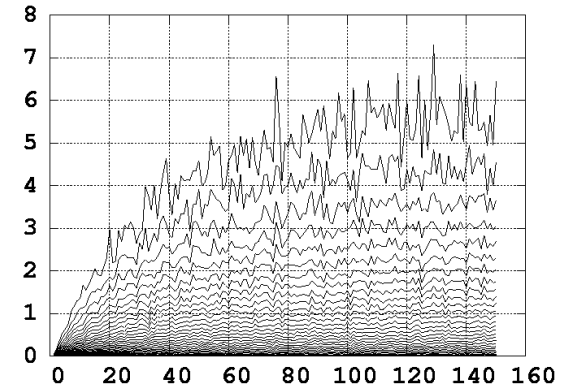
$$X_i = a \cdot \sin(\omega i) + \epsilon_i$$



(a) ARMA Process



(b) Periodic Process



(c) Exponential Process

Figure 4: Quantiles of time-dependent processes.

The cycle length of the sine oscillation is given by $T = \frac{2\pi}{\omega}$ with the amplitude a. We choose $T = 50$ and $a = 1$. Again $\{\epsilon_i\}_{i=1}^{\infty}$ is an independent Gaussian white noise process. The estimates of quantiles are depicted in Figure 4(b). The periodic behaviour is visible for every depicted quantile.

An Exponential Process

In the previous example we estimated quantiles of normally distributed processes. In this example we chose a process which is governed by an exponential distribution (see e.g. [Law and Kelton, 2000]). It is defined by

$$X_i = \epsilon'_i \cdot b(1 - e^{(i\frac{ln(0.05)}{l})}).$$

The process $\{\epsilon'_i\}_{i=1}^{\infty}$ is similar to the independent Gaussian white noise process, but its distribution is exponential with $\beta = 1$. The parameter b stretches the distribution. The part in brackets of the formula causes the process to slowly converge towards its marginal distribution. This is depicted in Figure 4(c). Both the convergence and the exponential character of the distribution is clearly apparent.

In general, the quantiles of areas with lower probability seem to fluctuate more than the ones of high probability. In Figure 4(a) and Figure 4(b) this can be observed when comparing the bounds 0 and 1 with the center (around 0.5) of the distribution. Because the distribution in Figure 4(c) is not symmetrical, the quantiles at bound 1 fluctuate more than the ones at bound 0. These examples show, that our approach of depicting quantiles is suitable for both symmetrical and asymmetrical distributions, as well as for converging and non converging processes. [Eickhoff et al., 2005] recommends the use of at least 50 independent replications to ensure a set of at least 5 different quantiles.

A Bounded Random Walk

The next process is based on a random walk X'_i, which is defined by

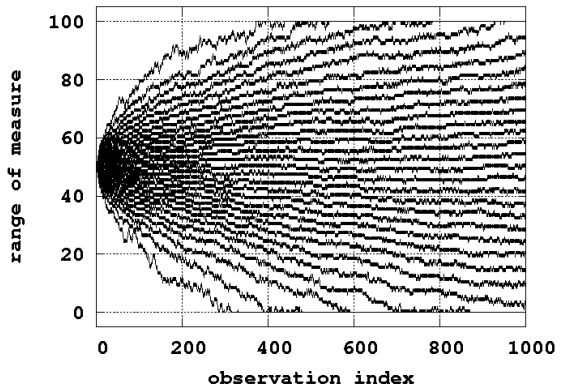
$$X'_i = \begin{cases} X'_{i-1} + 1, & \text{with probability } 0.5, \\ X'_{i-1} - 1, & \text{with probability } 0.5, \end{cases}$$

with the initial state $X'_0 = 50$. The process X'_i can take any value between $-\infty$ and $+\infty$. The final process X_i is bounded, so that its range is the interval $[0, 100]$:

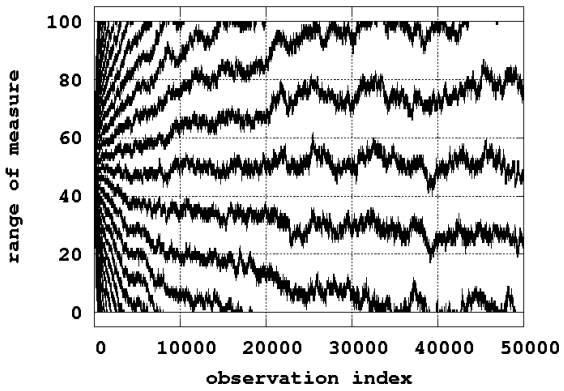
$$X_i = \begin{cases} 0, & \text{if } X'_i < 0, \\ X'_i, & \text{if } 0 \leq X'_i \leq 100, \\ 100, & \text{if } X'_i > 100. \end{cases}$$

A similar process was used in [Bause and Beilner, 1999]. Because X_i is bounded a marginal distribution for $i = \infty$ exists.

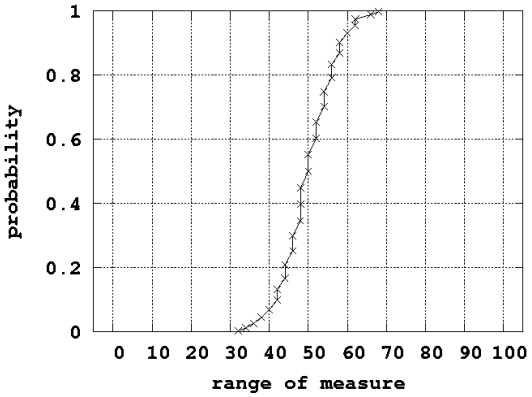
The peculiarity of this process is that the expected value $\mathrm{E}[X_i] = 50$ is constant over i, whereas all quantiles other than the median are not constant and converge to the thresholds 0 and 100, see Figure 5(a) and 5(b). $F_{X_i}(x)$ is very steep around $x = 50$ for small i, see Figure 5(c). After a long simulation time the shape of $F_{X_i}(x)$ is completely different. For large i it is very flat around $x = 50$, see Figure 5(d). However, the expected value $\mathrm{E}[X_i]$ is constant for all i. Analysis of mean values only would show a constant behaviour, even though this process is transient and the cumulative distribution is slowly converging to its marginal distribution.



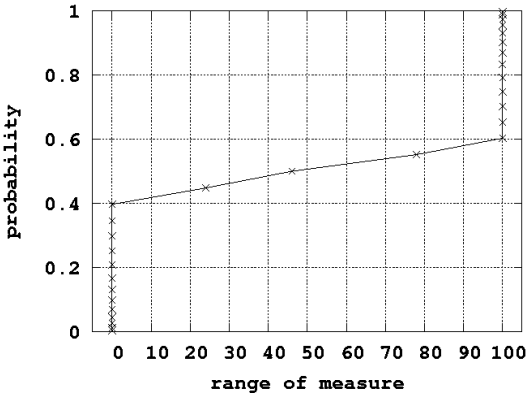
(a) short run behaviour of quantiles



(b) long run behaviour of quantiles



(c) $\hat{F}_{X_{50}}$: ECDF at index 50



(d) $\hat{F}_{X_{50000}}$: ECDF at index 50000

Figure 5: Quantiles and ECDFs of a bounded random walk.

A Periodic Queueing Model

A periodic behaviour can be introduced into a queueing model in two ways. On the one hand, the system arrivals could be governed by an oscillating function. On the other hand, the service process could be influenced by an oscillating function. In this example we choose a single server system with an unbounded queue. The interarrival process is governed by a Poisson distribution. The service process is deterministic and periodic. We denote this queueing process as $M/D_{periodic}/1/\infty$. The service time μ_i of the ith customer is defined by:

$$\mu_i = a \cdot \sin(\omega i) + \mu$$

The average service time μ is a positive value. a is the amplitude, with $0 \leq a \leq \mu$, to avoid negative values of the service time μ_i. The cycle length $T = \frac{2\pi}{\omega}$ of the sine oscillation is also a positive value.

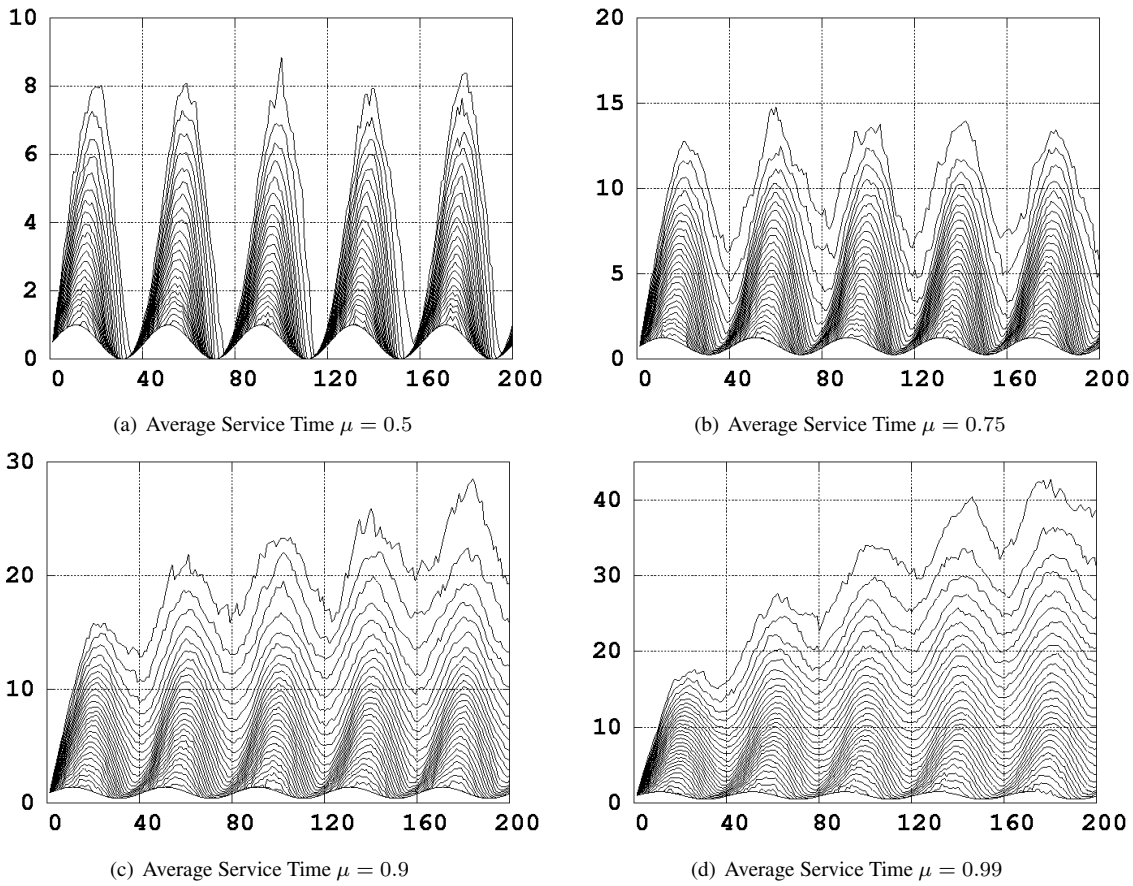
In our experiments we choose $\mu = \{0.5, 0.75, 0.9, 0.99\}$, $a = 0.5$, $T = 40$ and the average interarrival time is 1.0. We observed the response time, i.e. the time spend in queue plus the time spend in service, of consecutive customers. The results are depicted in Figure 6. (Note the different y-scale of each plot.) The periodic influence is clearly evident. Furthermore, the influence is different for each quantile. The peaks of higher quantiles are shifted by about $T/4$, whereas the peaks of lower quantiles stay close to the original periodic behaviour. Higher quantiles describe long queue length. Therefore, it can be assumed that a long queue damps the effect of the periodic behaviour. The peaks become higher and wider for an increasing μ so that they grow together. (Compare Figure 6(a) and Figure 6(d).)

A Chaotic Queueing Model

Chaotic systems are nonlinear, aperiodic and depend heavily on initial conditions. Usually they have a control parameter, which can cause the chaos to appear or disappear. The logistic equation

$$\mu_i = a\mu_{i-1}(1 - \mu_{i-1}) \tag{9}$$

shows chaotic behaviour if the initial state μ_0 is not a fixed point of Equation (9). This would lead to a constant μ_i. a is a positive constant with $0 < a \leq 4$. For some settings of a the process μ_i converges to one value. For other settings of a it jumps between a certain number of values after an initial phase. And for some settings of a the process μ_i shows no pattern at all. Small changes of a can lead to completely different behaviour of μ_i. For a detailed discussion on Equation (9) see [Sprott, 2003]. If the logistic equation is implicitly hidden in a model, it is very hard to get an insight into its behaviour by analyti-



(a) Average Service Time $\mu = 0.5$

(b) Average Service Time $\mu = 0.75$

(c) Average Service Time $\mu = 0.9$

(d) Average Service Time $\mu = 0.99$

Figure 6: Quantiles of the response time of the $M/D_{periodic}/1/\infty$ system.

cal methods. We choose the logistic equation to define the service time μ_i of the ith customer in a single server system. This explicitly introduces a chaotic behaviour and we incorporate it in the queueing model $M/D_{logistic}/1/\infty$. A process of this kind is analytically tractable only if the exact value of a is known.

In our experiments we observed the response time of consecutive customers. The average interarrival time of the Poisson process is 1.0. We set $a = \{2, 1 + \sqrt{8}, 1 + \sqrt{8} + 0.01, 4\}$ and $\mu_0 = 0.3$. For $a = 2$ (see Figure 7(a)) the queueing model shows a short warm up period. After this, μ_i is constant, and therefore, the estimated transient quantiles seem to be stable. The point $a = 1 + \sqrt{8}$ is the onset of a window, in which μ_i jumps between three values. This is depicted in Figure 7(c). Figure 7(b) does not show this behaviour, even though the value of a is very similar. For $a = 4$ the depiction of the quantiles does not show any pattern. Furthermore, the time evolution of higher quantiles is not always exactly similar to those of lower quantiles. For example between the 40th and the 45th customer in Figure 7(d) the lowest quantile is on a constant high level but higher quantiles are increasing.

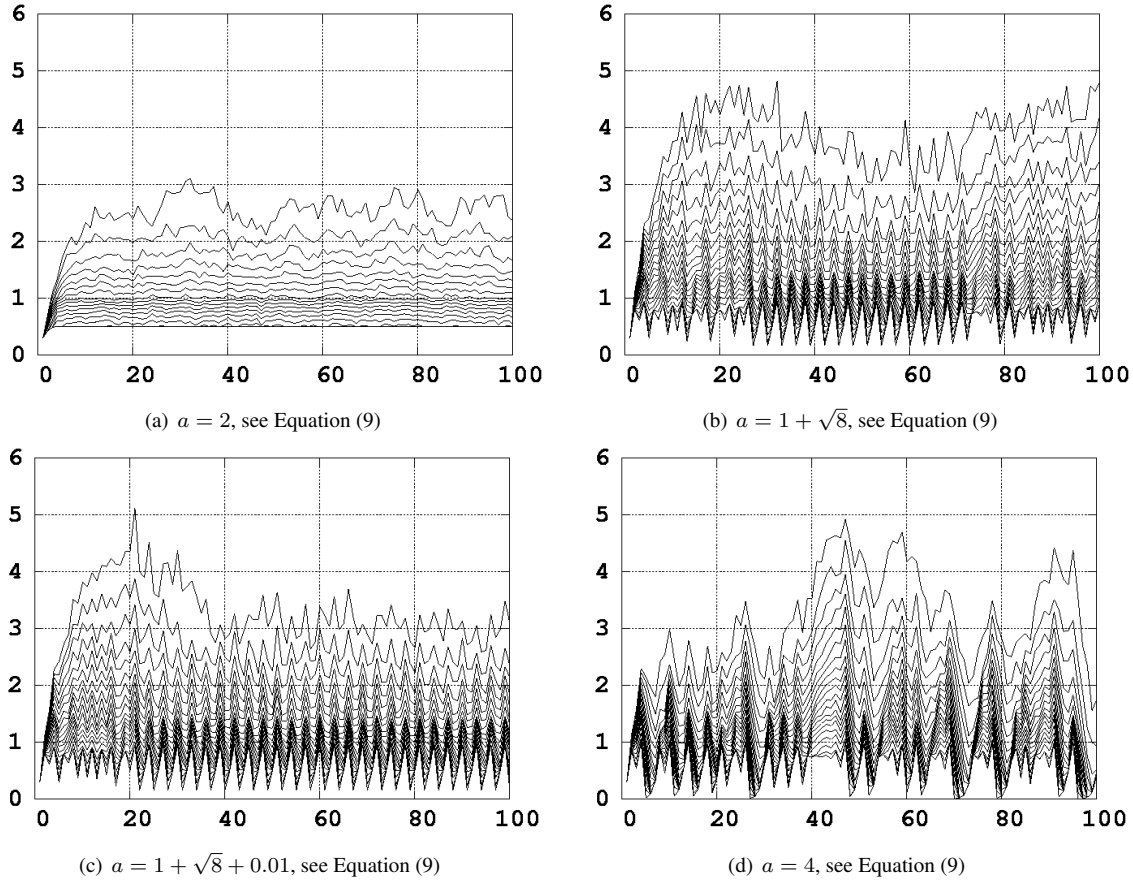
M/M/1/10 versus M/P/1/10

In the experiments described in this section we compare the $M/M/1/10$ queue with the $M/P/1/10$ queue. In the second queue the service process $X^{(P)}$ is governed by the Pareto distribution

$$F_{X^{(P)}}(x) = 1 - x^{-\alpha}, \quad x > 0.$$

To ensure that the first and the second moment of the Pareto distribution exists, we choose $\alpha = 3$. Therefore, $\mathrm{E}\left[X^{(P)}\right] = 1.5$ and $\mathrm{Var}\left[X^{(P)}\right] = 0.75$. To obtain comparable results, we choose the service process $X^{(M)}$ of the $M/M/1/10$ queue with the same expected value $\mathrm{E}\left[X^{(M)}\right] = 1.5$. The variance is in this case $\mathrm{Var}\left[X^{(M)}\right] = 2.25$. In both queueing models the average interarrival time is 1.0 and the maximum permitted queue length is nine customers plus one customer in service. Customers which arrive at a completely filled queue are rejected. Both queues are stable because their queue length is bounded.

We observed the response times in the two models for accepted customers. The results of our transient quantile estimation are shown in Figure 8. The quantiles converge to their steady state values. By comparing Figure 8(a) and Figure 8(c) it becomes obvious, that the probability distribution of the $M/P/1/10$ model is more centered around its expected value than the steady state distribution of the $M/M/1/10$ model.



(a) $a = 2$, see Equation (9)

(b) $a = 1 + \sqrt{8}$, see Equation (9)

(c) $a = 1 + \sqrt{8} + 0.01$, see Equation (9)

(d) $a = 4$, see Equation (9)

Figure 7: Quantiles of the response time of the $M/D_{logistic}/1/\infty$ system.

This is due to its smaller variance: $\mathrm{Var}\left[X^{(P)}\right] < \mathrm{Var}\left[X^{(M)}\right]$. The highest quantile of the $M/P/1/10$ model fluctuates more than the highest quantile of the $M/M/1/10$ model. Due to our choice of α, higher moments of the Pareto distribution do not exist, so this may cause the higher fluctuation of higher quantiles. In an additional experiment we started the replications with a completely filled queue. Theses results are plotted in Figure 8(b) and Figure 8(d). The 10 initial customers engender a non-monotonic convergence of the quantiles. For more information about quantile estimation of a $M/P/1$ model see [Fischer et al., 2001].

CONCLUSIONS

We have described two methods of selecting several quantiles with non-overlapping and non-disjoint confidence intervals. The first method operates in the rank domain, the second in the probability domain. Both methods delivered similar results. We recommend the second method based on Inequality (5) because of its lower complexity.

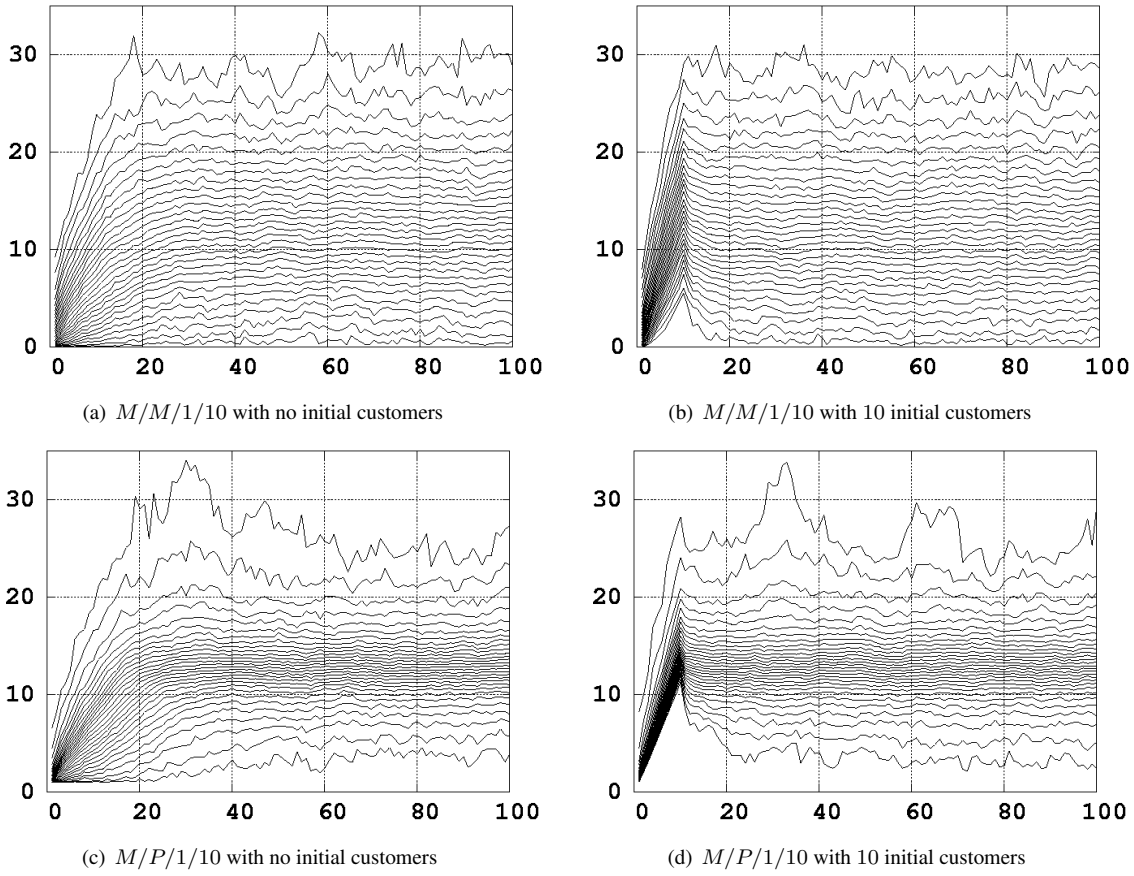
The use of multiple independent replications enables analysis of the evolution of several quantiles over time. Such analysis appears to be suitable for studying performance of a variety of different stationary, non-stationary and transient processes. Furthermore, this approach can be used in steady state simulation, as well as in finite-horizon simulation. In [Eickhoff et al., 2005] the use of at least 50 independent replications is recommended to make sure that the selected set of quantiles is reasonably large. In finite-horizon simulation the replications do not need to be processed in parallel. Therefore, a large number of replications, e.g. $p = 1000$, is feasible.

Our examples show that the analysis of quantiles can provide a deeper insight into the analyzed process than its mean value analysis. Drawing conclusions entirely based on mean value analysis is not recommended for complex models. If analytical investigations of a model fail, transient quantile estimation may be a good choice to obtain a deeper insight.

REFERENCES

[Avramidis and Wilson, 1995] Avramidis, A. N. and Wilson, J. R. (1995). Correlation-induction techniques for estimating quantiles in simulation experiments. *Proceedings of the 1995 Winter Simulation Conference*, pages 268–277. ISBN 0-7803-3018-8.

[Avramidis and Wilson, 1998] Avramidis, A. N. and Wilson, J. R. (1998). Correlation-induction techniques for estimating quantiles in simulation experiments. *Operations Research*, 46(4):574–591. ISSN 1526-5463.



(a) $M/M/1/10$ with no initial customers

(b) $M/M/1/10$ with 10 initial customers

(c) $M/P/1/10$ with no initial customers

(d) $M/P/1/10$ with 10 initial customers

Figure 8: Quantiles of the response time of the $M/M/1/10$ system in comparison with the $M/P/1/10$ system.

[Bause and Beilner, 1999] Bause, F. and Beilner, H. (1999). Intrinsic problems in simulation of logistic networks. *Proceedings of the 11th European Simulation Symposium and Exhibition (ESS99)*, pages 193–198. ISBN 1-56555-117-x.

[Bause and Eickhoff, 2003] Bause, F. and Eickhoff, M. (2003). Truncation point estimation using multiple replications in parallel. *Proceedings of the 2003 Winter Simulation Conference*, pages 414–421. ISBN 0-7803-8131-9.

[Chen, 2002] Chen, E. J. (2002). Two-phase quantile estimation. *Proceedings of the 2002 Winter Simulation Conference*, pages 447–455. ISBN 0-7803-7614-5.

[Chen and Kelton, 1999] Chen, E. J. and Kelton, W. D. (1999). Simulation-based estimation of quantiles. *Proceedings of the 1999 Winter Simulation Conference*, pages 428–434. ISBN 0-7803-8786-4.

[Chen and Kelton, 2001] Chen, E. J. and Kelton, W. D. (2001). Quantile and histogram estimation. *Proceedings of the 2001 Winter Simulation Conference*, pages 451–459. ISBN 0-7803-7307-3.

[Conover, 1999] Conover, W. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons, Inc., New York. ISBN 0-471-16068-7.

[Eickhoff et al., 2005] Eickhoff, M., McNickle, D., and Pawlikowski, K. (2005). Depiction of transient performance measures using quantile estimation. *Proceedings of the 19th European Conference on Modelling and Simulation (ECMS'2005)*, pages 358–363. ISBN 1-84233-112-4.

[Fischer et al., 2001] Fischer, M. J., Masi, D. M. B., Gross, D., Shortle, J., and Brill, P. H. (2001). Using quantile estimates in simulating internet queues with pareto service times. *Proceedings of the 2001 Winter Simulation Conference*, pages 477–485. ISBN 0-7803-7307-3.

[Fishman and Yarberry, 1997] Fishman, G. S. and Yarberry, L. S. (1997). An implementation of the batch means method. *INFORMS Journal on Computing*, 9(3):296–310. ISSN 1091-9856.

[Goldsman and Schmeiser, 1997] Goldsman, D. and Schmeiser, B. W. (1997). Computational efficiency of batching methods. *Proceedings of the 1997 Winter Simulation Conference*, pages 202–207. ISBN 0-7803-4278-X.

[Hamilton, 1994] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press. ISBN 0-691-04289-6.

[Hashem and Schmeiser, 1994] Hashem, S. and Schmeiser, B. W. (1994). Algorithm 727 quantile estimation using overlapping batch statistics. *ACM Transactions on Mathematical Software*, 20(1):100–102. ISSN 0098-3500.

[Heidelberger and Lewis, 1984] Heidelberger, P. and Lewis, P. (1984). Quantile estimation in dependent sequences. *Operations Research*, 32(1):185–209. ISSN 1526-5463.

[Heidelberger and Welch, 1981] Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence intervall generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245. ISSN 0001-0782.

[Igelhart, 1976] Igelhart, D. L. (1976). Simulating stable stochastic systems, vi: Quantile estimation. *Journal of the ACM*, 23(2):347–360. ISSN 0004-5411.

[Jain and Chlamtac, 1985] Jain, R. and Chlamtac, I. (1985). The P^2 algorithm for dynamic calculations of quantiles and histograms without storing observations. *Communications of the ACM*, 28(10):1076–1085. ISSN 0001-0782.

[Jin et al., 2003] Jin, X., Fu, M. C., and Xiong, X. (2003). Probabilistic error bounds for simulation quantile estimators. *Management Science*, 14(2):230–246. ISSN 0025-1909.

[Law and Kelton, 2000] Law, A. M. and Kelton, W. D. (2000). *Simulation Modeling and Analysis*. McGraw-Hill Higher Education, New York. ISBN 0070592926.

[L'Ecuyer et al., 2002] L'Ecuyer, P., Simard, R., Chen, E. J., and Kelton, W. D. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6):1073–1075. ISSN 1526-5463.

[Lee et al., 1999] Lee, J.-S. R., McNickle, D., and Pawlikowski, K. (1999). Quantile estimation in sequential steady-state simulation. *Proceedings of the 13th European Simulation Multiconference*, pages 168–174.

[Raatikainen, 1987] Raatikainen, K. E. E. (1987). Simultaneous estimation of several percentiles. *SIMULATION*, 49(4):159–164. ISSN 0037-5497.

[Raatikainen, 1990] Raatikainen, K. E. E. (1990). Sequential procedure for simultaneous estimation of several percentiles. *Transactions of the Society for Computer Simulation*, 7(1):21–44. ISSN 0740-6797.

[Seila, 1982] Seila, A. F. (1982). A batching approach to quantile estimation in regenerative simulations. *Management Science*, 28(5):573–581. ISSN 0025-1909.

[Sprott, 2003] Sprott, J. C., editor (2003). *Chaos and Time-Series Analysis*. Oxford University Press. ISBN 0198508395.

[Wood and Schmeiser, 1994] Wood, D. C. and Schmeiser, B. (1994). Consistency of overlapping batch variances. *Proceedings of the 1994 Winter Simulation Conference*, pages 316–319. ISBN 0-7803-2109-X.

AUTHOR BIOGRAPHIES

MIRKO EICKHOFF holds a Diploma degree in Computer Science from the University of Dortmund. His research interests are in the area of output analysis of discrete event simulation using multiple replications. His diploma thesis is part of the Collaborative Research Center "Modelling of Large Logistic Networks" (559) supported by the Deutsche Forschungsgemeinschaft. He worked for Delmia (Germany) in the area of workload balancing in manufacturing industry. In 2004 he received a Doctoral Scholarship of the University of Canterbury and is currently a Ph.D. candidate in Computer Science in the Simulation Research Group at this University. His e-mail address is m.eickhoff@cosc.canterbury.ac.nz.

DON MCNICKLE is an Associate Professor of Management Science in the Management Department at the University of Canterbury. His research interests include queueing theory, networks of queues and statistical aspects of stochastic simulation. He is a member of INFORMS and the Operational Research Society. His e-mail address is don.mcnickle@canterbury.ac.nz.

KRZYSZTOF PAWLIKOWSKI is a Professor in Computer Science at the University of Canterbury, in Christchurch, New Zealand. The author of over 130 research papers and four books; has given invited lectures at over 80 universities and research institutes in Asia, Australia, Europe and North America. Alexander-von-Humboldt Research Fellow (Germany) in 1983-84 and 1999. His research interests include performance modelling of telecommunication networks, discrete-event simulation and distributed processing. Senior Member of IEEE, member of ACM and SMSI. His e-mail address is krys.pawlikowski@canterbury.ac.nz.