COMMUN. STATIST. -SIMULA., 20(1), 375-390 (1991)

MODELING SERVICE–TIME DISTRIBUTIONS FOR QUEUEING NETWORK SIMULATION

Kimmo E. E. Raatikainen

University of Helsinki, Department of Computer Science
Teollisuuskatu 23–25, SF-00510 Helsinki, FINLAND

Key Words and Phrases: modeling of simulation input data; nonlinear data transformation; robust regression.

ABSTRACT

Shapes of service-time distributions in queueing network models have a great impact on the distribution of system response-times. It is essential for the analysis of response-time distribution that the modeled service-time distributions have the correct shape. Tradionally modeling of service-time distributions is based on a parametric approach by assuming a specific distribution and estimating its parameters. We introduce an alternative approach based on the principles of exploratory data analysis and nonparametric data modeling. The proposed method applies nonlinear data transformation and resistant curve fitting. The method can be used in cases, where the available data is a complete sample, a histogram, or the mean and a set of 5-10 quantiles. The reported results indicate that the proposed method is able to approximate the distribution of measured service times so that accurate estimates for quantiles of the response-time distribution are obtained.

1. INTRODUCTION

When queueing network models are simulated, the analyst must specify the service-time distributions and the branching probabilities. The traditional approach of modeling service-time distributions has been to assume that the distribution belongs to the Coxian family, and then to estimate the parameters. The Coxian family of distributions, introduced by Cox (1955), consists of all the distributions having a

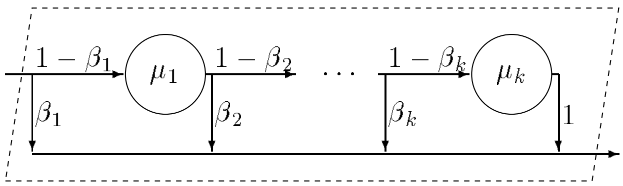

FIG 1. Phase diagram of k-stage Coxian distribution.

rational Laplace transform. Usually a Coxian distribution is characterized as a phase model (FIG 1). The Coxian assumption is reasonable, since Whitt (1974) showed that any non-negative random variable can be approximated arbitrarily closely by a Coxian distribution. However, the fundamental problem of this parametric approach is the identification of the Coxian distribution.

The simplest (and most commonly used) method to identify the Coxian distribution is to assume a certain number of stages – usually the number of stages has been one or two – and then to solve the parameters using the method of moments. However, we will show that the method of moments leads to inadequate models, when the response-time distribution is analyzed. This is not surprising, since it was empirically shown by Pearson, Johnson and Burr (1979) that the quantiles may be quite different although the considered distributions have the same first four moments. Moreover, the comparisons given by Khoshgoftaar and Perros (1987) indicate that the method of moments is a poor choice to identify the service-time distribution, if the moments are not known but must be estimated. Several identification techniques that are more sophisticated have been proposed in the literature: e.g. Bux and Herzog (1977), Lazowska (1977), and Lazowska and Addison (1979). None of these methods is without some problematic features. In view of the wide flexibility of Coxian distributions, the situation is not unexpected.

In this paper we propose an alternative approach of modeling service-time distributions. Our approach is based on the principles of exploratory data analysis and nonparametric data modeling. These principles are widely discussed by Mosteller and Tukey (1977), by Parzen (1979), and by Hoaglin, Mosteller and Tukey (1985). The basic idea is first to transform the data and then to apply resistant curve fitting in order to obtain the quantile function.

Regarding available identificaton data we consider three different cases. All of them are treated almost identically. In the first case we assume that we have a complete measurement sample available. In computer capacity planning, for example, the usage of complete measurement samples over long periods of time is impractical. Therefore we consider, as the second case, the situation where the measurement data is reducted to a histogram. Since typical measurement data of computer resource usage has a skew orientation, ordinary histograms having equal bin widths are inadequate. Instead, we use "octaves": that is a $\log_2$-based classification introduced by Tukey (1977). The third case is primarily intended to the case of predicted workload. We assume that we know the estimates or predictions of the following indices: the mean and a set of 5-10 percentiles. In this study we have used the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, and 99^{th} percentiles.

We begin this paper by discussing two important background issues. Firstly, using two simple example networks, we show that the shape of the service-time distribution – not only its first two moments – has an important effect on the distribution of response times. Secondly, we introduce a technique to analyze the shape of a distribution.

In Section 3, we introduce our method of modeling service-time distributions. The method is based on fitting a continuous piecewise Hermite polynomial to the transformed data by iteratively reweighted least-squares or IRLS (Holland and Welsch 1977). In Section 4, we present our experimental results; the effects that the service-time distribution has upon the response-time distribution are analyzed in an open single-server queue.

2. BACKGROUND CONSIDERATIONS

We consider two important issues that according to our experience have been widely ignored. We discuss, as the first issue, the effects that service-time distributions have upon the response-time distribution. Since analytical results are not available, we use two simple examples to demonstrate some of the possible effects. The second issue is related to data analysis. In simulation textbooks, the "standard" distributions of mathematical statistics – such as the Gaussian, exponential, gamma, Weibull, and beta distribution – have received most attention. Empirical distributions are usually covered only briefly. However, the distributions of the "real-world" are typically quite different than those "ideal" distributions. Therefore we discuss, as the second issue, the analysis of the shape of a distribution.

2.1 Impacts on Response Time Distributions

In Example 1, we analyzed a simple queueing network model (FIG 2). In order to show the significance of the service-time distribution, we simulated the model with three different service-time distributions in the CPU node. All the three distributions – labeled as A, B, and C – were two-stage Coxians, and they had equal means and variances. The characteristics of the distributions are given in Table I.

With each service-time distribution, we generated 100 independent sequences of response times. The length of each sequence was 10,000. Simulations were initialized with the steady-state load distribution and a warm-up period of 1,000 response times. The means of estimated response-time indices are given in Table II, together with their 95% confidence limits. The confidence limits are based on the normal approximation. Although the simulated means are practically equal, the differences between percentiles are significant.

Node	Type	Demand
Terminals	IS	30
I/O	IS	0.5
CPU	PS	1
Network Characteristics		
Population		40
CPU-utilization		.983
Mean response time		10.7

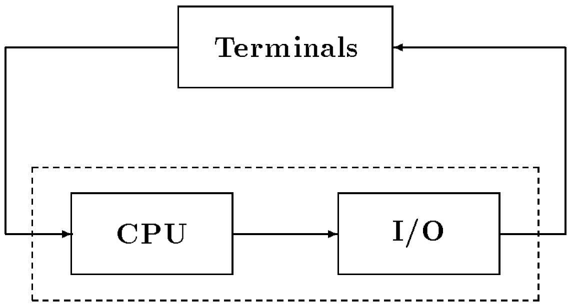

FIG 2. Queueing network model of Example 1.

TABLE I
Summary of service-time distributions in Example 1

	A	B	C
β_1	0	0	0
β_2	.990	.985	.980
μ_1	2.7584	4.5900	10.469
μ_2	.015687	.019178	.022112
Mean	1	1	1
CV	9	9	9
Skewness	21.1	17.2	14.9
Kurtosis	598	400	300

TABLE II
Summary of response-time distributions in Example 1

	A	B	C
Mean	10.690±.586	10.901±.639	10.484±.541
Ptile			
.05	.440±.015	.338±.012	.218±.006
.10	.687±.024	.519±.018	.331±.009
.25	1.375±.051	1.003±.036	.622±.016
.50	2.852±.111	1.956±.073	1.150±.030
.75	5.654±.217	3.684±.139	2.000±.053
.90	9.943±.357	6.384±.238	3.243±.090
.95	13.788±.476	8.976±.334	4.477±.130

Population	20
Service Demands	
CPU	1.033
IO	.356
TTY	30
System	
Throughput	.630
Mean Response	2.829
Mean Load	1.724
Utilizations	
CPU	.630
IO	.217

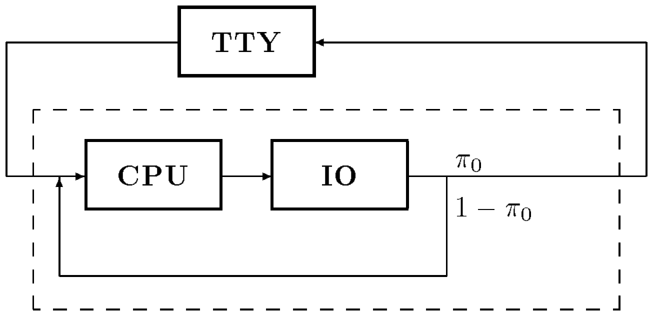

FIG 3. Queueing network model of Example 2.

This simple example makes two important phenomena clearly visible. First, the mean response-time characterizes the performance incompletely. The second point is that the mean and variance do not provide enough information to determine the service-time distribution, when the response-time distribution is analyzed.

In Example 2, we consider another simple closed queueing network model (FIG 3). This example has two objectives. Firstly, we want to compare exponentially distributed service times to service times whose distribution resembles more accurately the measured cpu times. Secondly, we consider what kinds of effects the joint distribution of service demands can have on the response times. The model consists of three nodes: CPU, TTY, and IO. The queueing disciplines are processor sharing (PS), no queueing or infinite server (IS), and first-come-first-served (FCFS), respectively. The service-demand distribution in the IO node is exponential with mean 0.356. The "thinking" time in the TTY node is exponentially distributed with mean 30. Two different service-demand distributions are used in the CPU node: an exponential with mean 1.033 and a three-stage Coxian, $\mu_1 = 2480.$, $\mu_2 = 5.320$, $\mu_3 = .07244$, $\beta_1 = \beta_2 = 0$, $\beta_3 = .93878$ (mean service-demand is 1.033).

We generated 100 independent sequences of the system response-times with both the service-demand distributions. Each sequence contained 10,000 observations. The simulations were initialized using the steady-state load-distribution. A warm-up period of random length $[1024, 2047]$ was used to reduce the initialization bias. Five different probabilities of cycle completion were used: $\pi_0 = 1, {}^1/_2, {}^1/_4, {}^1/_8, {}^1/_{16}$.

The boxplots of response-time distributions are in FIG 4. The given indices are the means (•) and the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles of the response times. In the given form of boxplot, the ends of box denote the lower and upper quadrant. The line inside the box denote the median. The "whiskers" reach the 10^{th} and 90^{th} percentile. Separate lines are used to denote the 5^{th}, 95^{th}, and 99^{th} percentile.

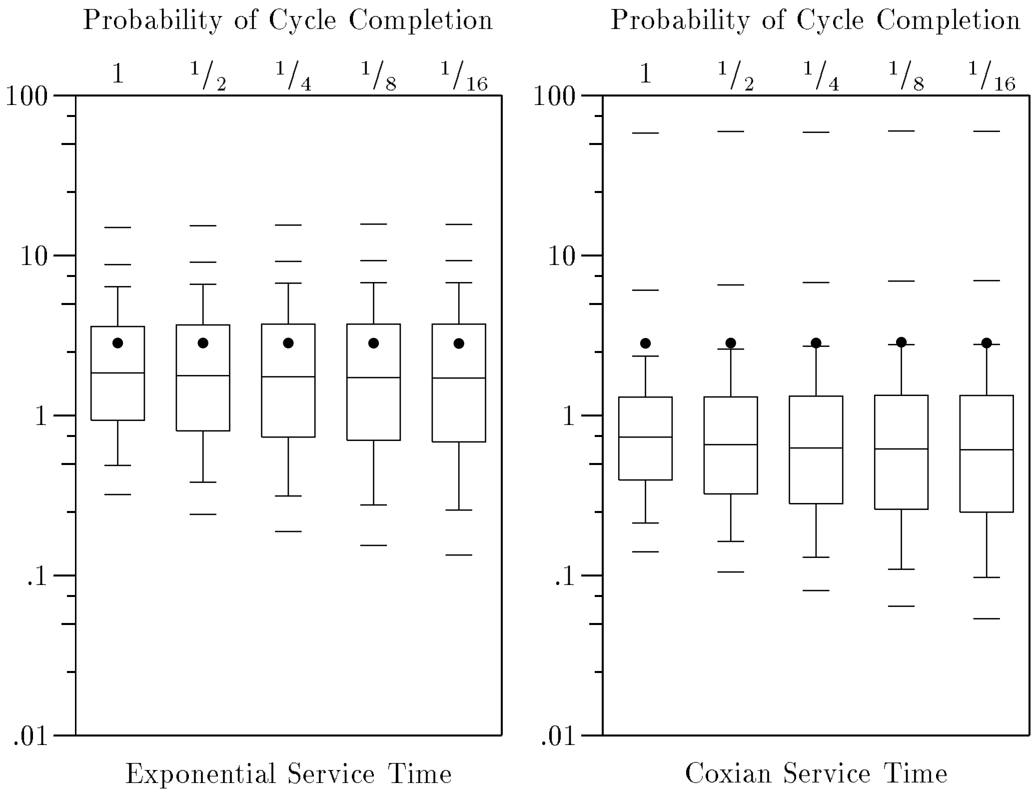

FIG 4. Summaries of response times in Example 2.

The differences between the exponential and Coxian service-time distribution clearly indicate the significance of the service-demand distribution, when the quantiles of response times are considered. In this example, the joint distribution has a significant effect only on the lower quantiles, approximatively up to the lower quadrant. The median is only slightly affected, and the upper quantiles are practically unaffected.

2.2 Analyzing the Shape of Distribution

The two examples above clearly show that the mean and variance provides an insufficient characterization of the service-time distribution when the response-time distribution is analyzed. Therefore, we must consider other aspects of the distribution as well. Although it is common practice to use higher moments to characterize skewness, kurtosis, etc.; we prefer techniques based on order statistics.

Theoretical and practical motivations for preference of quantiles to moments in modeling the distributions are given in Hoaglin, Mosteller, and Tukey (1985) and in Mosteller and Tukey (1977). Our preference is based on the following two reasons. Firstly, it is evident that most people understand the quantiles much better than the moments, perhaps with the exception of the mean. Secondly, most quantiles have better sampling properties than moments.

The shape of a distribution is not, in general, a precisely established concept. Therefore, the shape of an arbitrary distribution can not be correctly summarized by any specific measure; see the discussion in Chapter 1 of Mosteller and Tukey (1977). Instead, the best we can do is to compare the shape of an arbitrary distribution to the shape of a specific reference distribution. Our technique of summarizing the shape of a distribution is based on similar ideas as in Chapter 10 of Hoaglin, Mosteller and Tukey (1985), where the concept of *pseudosigma* was introduced. Pseudosigmas are intended to compare the shape of data to the Gaussian shape. The p^{th} pseudosigma is an estimator of Gaussian scale-parameter σ based on the p^{th} and $(1-p)^{\mathrm{th}}$ data quantiles.

The exponential distribution has the same central rôle in queueing network modeling as the Gaussian distribution has in classical statistics. The exponential distribution has a single parameter, scale $1/\lambda$. We define the concept of *pseudoscale* as an estimator of exponential scale based on a single data quantile. The reason to use only one quantile, not two as in pseudosigma, is that unlike the Gaussian distribution the exponential distribution is not symmeric. Formally, the p^{th} *pseudoscale*, $1/\lambda_p$, is defined as

$$1/\lambda_p = x_p/z_p \; , \tag{1}$$

where x_p is the p^{th} quantile of the data and z_p is the p^{th} quantile of the unit exponential distribution: $z_p = -\log(1-p)$.

The comparison of the data shape to the exponential shape is efficiently visualized, when we plot the logs of pseudoscales against the corresponding quantiles of the unit exponential. In order to demonstrate the pseudoscales we analyzed the five distributions considered in the previous two examples, six hyperexponential (equal-load) distributions, and six Erlang distributions. The analyzed hyperexponential and Erlang distributions have unit mean. The variances of the hyperexponential distributions were 2, 4, 8, 16, 32, and 64; and of the Erlang distributions 1/2, …, 1/64. The pseudoscales (in log-scale) of those distributions are plotted against the unit exponential in FIG 5.

3. MODEL FOR SERVICE-TIME DISTRIBUTION

In nonparametric data modeling the most widely used representations of the distribution function are splines and Fourier series; see e.g. Bean and Tsokos (1980). The Fourier series representation cannot be identified if we know only the mean and 5-10 quantiles. The splines might be used, but we have found that piecewise Hermite polynomials are better than splines. In data analysis, it is strongly recommended that the data should be transformed before the distribution is modeled, see e.g. Chapter 11 of Hoaglin, Mosteller and Tukey (1985) and Wold (1974). We use the pseudoscales introduced in the previous section as the initial data transformation.

Our method of modeling a service-time distribution is based on robust curve fitting to the logarithms of the pseudoscales. Therefore, our method is actually a standard technique of statistics: When the distribution does not belong to the

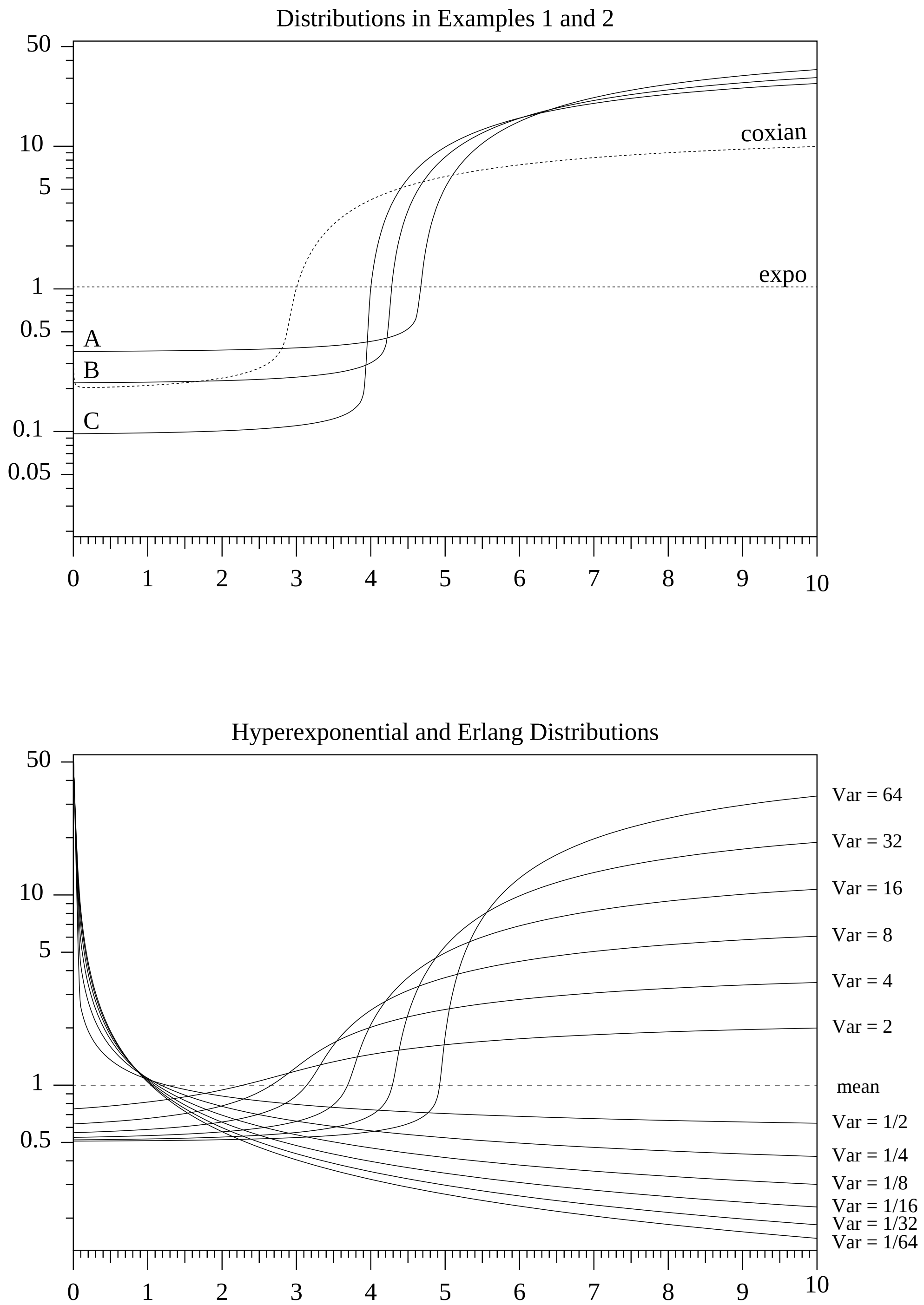

FIG 5. Pseudoscales of examined distributions.

"standard set" of distributions, the distribution is modeled as a monotonic function of a standard random variable. The best known of these frequency curves are the Pearson and Johnson curves, summarized for example in Chapter 12.4 of Johnson and Kotz (1970).

In statistics, the Gaussian distribution is the usual standard distribution. We, however, have chosen the unit exponential as our standard distribution. The most important reason is that the service-time distributions are non-negative. In addition, we have found that the logarithm of pseudoscales is usually a quite smooth function of the unit exponential.

In this section we describe our model and introduce the numerical methods of identifying it. The three cases of available data – complete sample, histogram, and the mean and a few percentiles – are modeled in the same way. The only difference is in the knot selection and tail-area modeling. When we know only the mean and a few percentiles, we must adopt a different procedure to identify the tail-area behaviour. In the two other cases, the tail-area modeling does not need any special treatment.

3.1 Model Specification

Our model is based on robust curve fitting to the transformed data. We model service times as a monotonic function of unit exponential. In practice, the service time is never infinite, but has always a finite upper limit. Therefore the modeled distribution has a discrete component at that limiting point.

Now we can formally specify our distribution model. The p^{th} quantile of the service time, x_p, is given by

$$x_p = \begin{cases} z_p \exp[\phi(z_p)] & , \text{if} \quad p \leq p_{max} \\ x_{max} & , \text{if} \quad p \geq p_{max} \end{cases} \quad , \tag{2}$$

where $z_p = -\log(1-p)$, x_{max} is the upper limit of service times, and $p_{max} = \Pr\{X < x_{max}\}$. From the available data we can calculate a set of values for $\phi(z)$:

$$\phi(z_i) = \log(x_i/z_i) \, , i = 1, \ldots, n \, , \tag{3}$$

where $z_i = -\log(1-p_i)$, and $p_i = \Pr\{X \leq x_i\}$.

The amount of available data varies considerably in the three cases. Complete samples contain hundreds, thousands, or even tens of thousands of observations. From the histogram data we can calculate typically 10-30 pseudoscales. In the case of percentile data the number of pseudoscales that can be determined is usually less than 10.

We use a piecewise cubic Hermite polynomial to model the $\phi(z)$-function:

$$\hat{\phi}(z) = \phi_{j-1}(1 - 3t^2 + 2t^3) + \phi_j(3t^2 - 2t^3) + \phi'_{j-1}h_j t(1-t)^2 - \phi'_j h_j t^2(1-t) \, , \tag{4}$$

when $z_{j-1} \leq z \leq z_j$, $j = 1, \ldots, m$; the z_j's are the knots; $h_j = z_j - z_{j-1}$; and $t = (z - z_{j-1})/h_j$.

The parameters of the model are the number and locations of knots, and coefficients ϕ_j and ϕ'_j. The piecewise cubic Hermite polynomial and its first derivative are continuous. The coefficients are the values of the polynomial and its derivative at the knots. The cubic spline has, in fact, the same form and parameters as the cubic Hermite polynomial. The only difference is that the spline has fewer free parameters: The coefficients ϕ'_j, $j = 1, \ldots, m$, are determined internally so that the second derivative of the spline is also continuous.

We prefer Hermite polynomials to splines for the following two reasons. Firstly, we do not necessarily need the continuity of the second derivative. Actually, we have found that the splines have many times abnormal bows, since the second derivative is forced to be continuous. Secondly, the function $z \exp[\phi(z)]$ must be nondecreasing. In the case of Hermite polynomial it is easy to modify, if necessary, the function $\hat{\phi}(z)$ so that $z\hat{\phi}'(z) \geq -1$, $\forall\, z$.

3.2 Model Identification

When the knots are determined and the pseudoscales are calculated from the data, we determine the coefficients by minimizing the weighted sum of squares:

$$\min \sum_{i=1}^{n} w_i \left(\phi(z_i) - \hat{\phi}(z_i) \right)^2 . \tag{5}$$

In order to obtain a robust and resistant curve, we use the method of iteratively reweighted least-squares with the Tukey-window; see Holland and Welsch (1977), and Chapter 14H of Mosteller and Tukey (1977). The minimization (5) is a sparse problem; each observation affects only four of the coefficients. Therefore, the problem can be solved efficiently by the LSQR-algorithm (Paige and Saunders 1982), even if we have tens of thousands of observations.

To complete the model identification we discuss the selection of knots. Some authors have suggested that the number and locations of knots should also be selected automatically. Then the problem will be nonlinear and difficult to solve. More often the knots are selected by rules of thumb, and trial and error; see Wegman and Wright (1983), Wold (1974), and Smith (1979), among others.

In the case of a complete sample we have used the 25^{th}, 50^{th}, 90^{th}, and 99^{th} quantiles of the unit exponential as the interior knots. The external knots are 0 and $z_{max} = -\log(1 - p_{max})$; and p_{max} is taken as $1 - 2/(3n + 1)$, where n is the sample size. Since we use robust regression the knot selection has only minor importance. However, with other sets of data these knots may not be good. Therefore we suggest that the fit should be examined graphically before the model is used. If large residuals are found, the number of knots may be increased and the positions changed. One obvious alternative is to select a set of feasible knots, for example 10-20 equally spaced values of z; and then proceed as described below.

In the case of histogram data we have typically about 20 observations. This implies that at most 3-5 internal knots should be used, see e.g. Wold (1974). The external knots are 0 and $z_{max} = -\log(1 - p_{max})$. Now $p_{max} = 1 - 2/(3N + 1)$, where N is the number of observations, on which the histogram is based. If N is not known, we must choose a suitable value, for example 10,000–100,000.

When the knots are at the data points, as generally recommended, the total number of possible sets of knots is moderate. We propose that the problem is solved with several sets of knots and that the best one is chosen. The selection of "best fit" is a common problem in statistics. Several criteria have been suggested in the statistical literature. In regression context the best known are Mallows' C_p (Mallows 1973) and Allen's PRESS (Allen 1974), while in time-series analysis the most widely used criterion is Akaike's AIC (Akaike 1974). All these criteria measure the goodness-of-fit by residuals and penalize the increase of the number of free parameters.

In our case the regression – that is the identification of $\hat{\phi}$-function – is only a tool to model the quantile function. Therefore we do not use residulas as a measure for the goodness-of-fit. Instead, we use the mean absolute probability distance,

$$\int \left|\tilde{F}(x) - \hat{F}(x)\right| d\tilde{F}(x) ,$$

where $\tilde{F}(x)$ is the measured and $\hat{F}(x)$ is the modeled cumulative distribution function. We penalize the increase of the number of parameters by multiplying the mean absolute probability distance by factor

$$\sqrt{\frac{n + 2m}{n - 2m}} ,$$

where n is the number of observations and m is the number of internal knots. Our multiplication factor is the same as in the FPE-criterion (Akaike 1969) that is used in identifying the order of the autoregressive model in time-series analysis.

When the available data is a set of percentiles and the mean, we first fit the curve to the available range of data. In the case of eight percentiles we can have only one or two internal knots. The external knots are 0 and z_n corresponding to the highest known percentile. As in the case of histrogram data, we try all of the feasible combinations of knots and select the best one. After that we adjust the tail-area behaviour so that the model provides the given mean.

In the area of the probability range $[p_n, p_{max}]$, where p_n is the highest known percentile and p_{max} is the known or approximated extremal percentile, we also use the Hermite model. The values of $\phi(z_n)$ and $\phi'(z_n)$ are already determied. So we need to determine the values of $\phi(z_{max})$ and $\phi'(z_{max})$.

We propose two alternative methods of determining $\phi(z_{max})$ and $\phi'(z_{max})$. The choice depends on whether or not we know the upper limit x_{max}. If x_{max} is known, we choose $\phi(z_{max}) = \log(x_{max}/z_{max})$ and determine $\phi'(z_{max})$ numerically so that the model provides the given mean. In the case of x_{max} being unknown, we choose

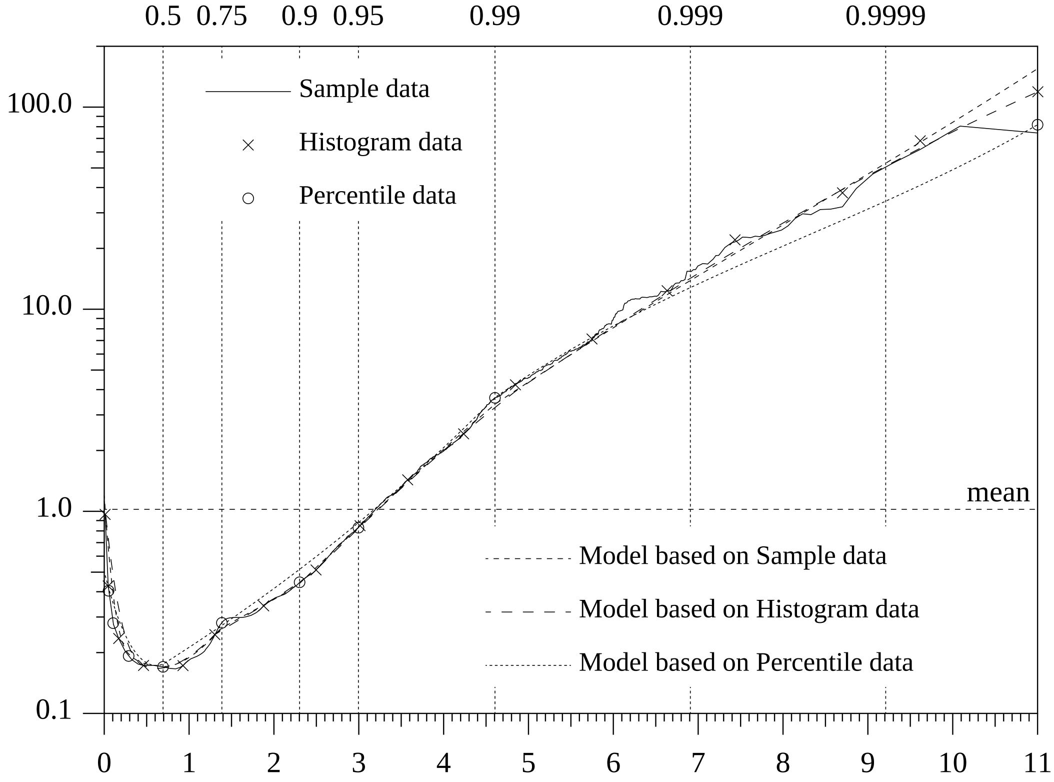

FIG 6. Pseudoscales for measured and modeled CPU demands.

$\phi'(z_{max}) = 0$ and determine $\phi(z_{max})$ numerically so that the model provides the given mean. The mean of the modeled distribution is obtained as

$$\bar{x} = \int_0^{z_{max}} z e^{\phi(z)-z} \, dz + z_{max} e^{\phi(z_{max})-z_{max}} \, ,$$

where $z_{max} = -\log(1 - p_{max})$.

3.3 Accuracy of Model Identification

The accuracy of the proposed identification method was examined using measured cpu-times in a VAX8600-system used in academic research. The measurement data contained the cpu-times of 44,052 interactive tasks executed during one week, daily from 9 a.m. to 6 p.m. In the first case we used the complete measurement data in model identification. In the second case we reduced the data to a histrogram having class boundaries 0.01, 0.02, 0.04, 0.08, ..., 1310.72 (seconds). In the third case we used the mean and 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, and 99^{th} percentiles. In the tail-area adjustment we used $z_{max} = 11$ and $x_{max} = 900$.

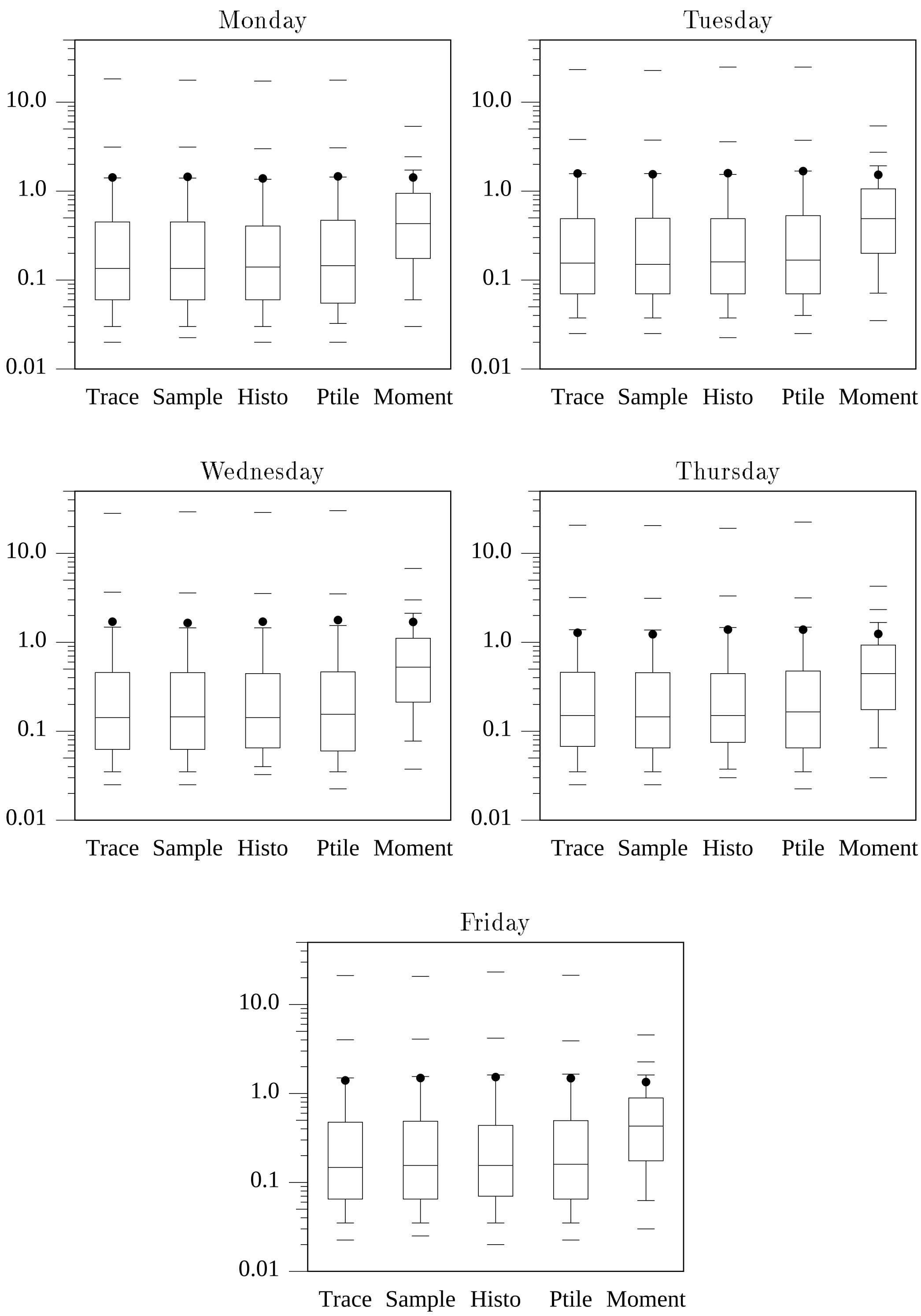

FIG 7. Boxplots of response times in trace-driven simulation.

The measured and modeled pseudoscales for the three cases are given in Figure 6. The figure indicates that the models based on sample and histogram data are close to each other and to the sample data. The great similarity of those two models is important: *We obtain great accuracy, even if we use only histogram data in the model identification.*

The model based on percentile data differs to a larger degree from the sample data than the first two models. However, we must remember that if we use predicted quantiles we do not know the behaviour between those quantiles. According to our experiences of modeling several other data sets of computer resource usages, the accuracy of the models reported in Figure 6 is typical.

4. EXPERIMENTAL RESULTS

We use an example to demonstrate the accuracy of our method. Since our ultimate objective is the distribution of response times, we analyzed how different techniques of modeling the service-time distribution affect the response-time distribution. We carried out simple trace simulations based on daily measurement data. The analyzed measurement data contains resource usages of all interactive tasks executed during one week; daily from 9 a.m. to 6 p.m. The response times were analyzed separately for each day from 10 a.m. to 5 p.m. Using the actual arrival times and ranks of the actual cpu-times we simulated an open CPU-queue assuming that the scheduling policy is processor sharing.

We studied five different service time distributions:

1. empirical;
2. the proposed model based on the whole sample;
3. the proposed model based on the histogram;
4. the proposed model based on the mean and eight percentiles: 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, and 99^{th} (The tail-area adjustment were based on the sample size, $p_{max} = 1-2/(3n+1)$ and observed maximum value of the service times.);
5. the hyperexponential distribution having the same first three moments as the empirical.

The service time of a specific task was taken as the quantile corresponding the rank. (The i^{th} rank corresponds to the quantile $(3i-1)/(3n+1)$, where n is the sample size.) The ties were broken by randomization.

The results are summarized in Figure 7 containing the boxplots of the obtained response-time distributions. The conclusions are self-evident. The proposed method, with all the three cases of identification data, provides accurate estimates of the response times. On the contrary, the hyperexponential assumption and the method of moments identifies a completely inappropriate shape of the service-time distribution: only one quantile of response times (out of 40) is not more than 10 % away from that of the trace.

SUMMARY

We have described a method of modeling the service-time distributions to be used in queueing network simulation. The method operates on transformed identification data that can be a complete sample, a histogram, or the mean and 5-10 percentiles. The service times are modeled as a monotonic function of the unit exponential random-variable. A piecewise cubic Hermite polynomial is fitted to transformed data. The coefficients of the polynomial are determined by robust regression based on the method of iteratively reweighted least-squares and Tukey-window. In addition, we have proposed a technique to select the knots of piecewise polynomial automatically.

The accuracy of the proposed method was examined using measured cpu-times of interactive tasks in a VAX8600. In addition to the accuracy of model identification, we analyzed how different methods of modeling the service-time distribution affect the obtained response-time distribution. The reported results indicate that the proposed method, with all the three classes of identification data, is able to approximate the shape of measured service times so that we obtain accurate estimates for quantiles of the response-time distribution.

ACKNOWLEDGEMENTS

This work was supported by the Emil Aaltonen Foundation.

BIBLIOGRAPHY

Akaike, H. (1969). 'Fitting Autogreressive Models for Prediction,' *Annals of Institute for Statistical Mathematics, 21,* 243–247.

Akaike, H. (1974). 'A New Look at the Statistical Model Identification,' *IEEE Transactions on Automatic Control, AC-19,* 716–723.

Allen, D. M. (1974). 'The Relatioship Between Variable Selection and Data Augmentation and a Method for Prediction,' *Technometrics, 16,* 125–127.

Bean, S. J. and Tsokos, C. P. (1980). 'Developments in Nonparametric Density Estimation,' *International Statistical Review, 48,* 267–287.

Bux, W. and Herzog, U. (1977). 'The Phase Concept: Approximation of Measured Data and Performance Analysis,' *Computer Performance* (K. M. Chandy and M. Reiser, Eds.). Amsterdam: North-Holland, 23–38.

Cox, D. R. (1955). 'A Use of Complex Probabilities in the Theory of Stochastic Processes,' *Proceedings of the Cambridge Philosophical Society, 51,* 313–319.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (Eds.) (1985). *Exploring Data Tables, Trends, and Shapes.* New York: John Wiley & Sons.

Holland, P. W. and Welsch, R. E. (1977). 'Robust Regression Using Iteratively Reweighted Least-Squares,' *Communications in Statistics – Theory and Methods, A6,* 813–827.

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions – 1.* Boston: Houghton Mifflin.

Khoshgoftaar, T. M. and Perros, H. G. (1987). 'A Comparision of Three Methods of Estimation for Approximating General Distributions by a Coxian Distribution,' *Modelling Techniques and Performance Evaluation* (G. Pujolle and S. Fdida, Eds.). Amsterdam: North-Holland, 81–89.

Lazowska, E. D. (1977). 'The Use of Percentiles in Modeling CPU Service Time Distributions,' *Computer Performance* (K. M. Chandy and M. Reiser, Eds.). Amsterdam: North-Holland, 53–66.

Lazowska E. D. and Addison, C. A. (1979). 'Selecting Parameter Values for Servers of the Phase Type,' *Performance of Computer Systems* (M. Arato, A. Butrimenko, and E. Gelenbe, Eds.). Amsterdam: North-Holland, 407–420.

Mallows, C. L. (1973). 'Some Comments on C_p,' *Technometrics, 15,* 661-675.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression. A Second Course in Statistics.* Reading: Addison-Wesley.

Paige, C. C. and Saunders, M. A. (1982). 'LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares,' *ACM Transactions on Mathematical Software, 8,* 43–71.

Parzen, E. (1979). 'Nonparametric Statistical Data Modeling,' *Journal of the American Statistical Association, 74,* 105–131.

Pearson, E. S., Johnson, N. L. and Burr, I. W. (1979). 'Comparisions of Percentage Points of Distributions with the Same First Four Moments, Choosen from Eight Different Systems of Frequency Curves,' *Communications in Statistics – Simulation and Computation, B8,* 191–229.

Smith, P. (1979). 'Splines as a Useful and Convenient Statistical Tool,' *The American Statistics, 33,* 49–69.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading: Addison-Wesley.

Wegman, E. J. and Wright, I. W. (1983). 'Splines in Statistics,' *Journal of the American Statistical Association, 78,* 351–365.

Whitt, W. (1974). 'The Continuity of Queues,' *Advances in Applied Probability, 6,* 175–183.

Wold, S. (1974). 'Spline Functions in Data Analysis,' *Technometrics, 16,* 1–11.

Received August 1989; Revised November 1990.