IIE Transactions (2011) **43**, 176–191

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2010.521806

Simulation-based estimation of cycle time using quantile regression

NAN CHEN[1] and SHIYU ZHOU[2,*]

[1]*Department of Industrial and Systems Engineering, National University of Singapore, Singapore*
[2]*Department of Industrial and Systems Engineering, 3254 Mechanical Engineering Bldg, University of Wisconsin–Madison, WI 53706, USA*
E-mail: szhou@engr.wisc.edu

Received July 2009 and accepted July 2010

Production cycle time is an important performance measure in manufacturing systems, and thus it is of interest to characterize distributional properties, such as quantiles, for informative decision making. This article proposes a non-linear quantile regression model for the relationship between stationary cycle time quantiles and corresponding throughput rates of a manufacturing system. The statistical properties of the estimated cycle time quantiles are investigated and the impact of dependent data from simulation output on parameter estimations is analyzed. Extensive numerical studies are presented to demonstrate the effectiveness of the proposed methods.

Keywords: Cycle time estimation, quantile regression, system throughput

1. Introduction

Production cycle time is usually represented by a random variable that characterizes the time required for a job or order to traverse through a system in a designed routine (see, for example, Hopp and Spearman (1996)). It is known to be an important performance measure of the responsiveness of a manufacturing system. Furthermore, cycle time is also strongly related to many other important aspects in a manufacturing system. For example, in a make-to-order production environment, delivery lead times, as well as order promising abilities, are to a large extent influenced by the production cycle time (Gordon, 1993). As another example, the famous Little's law indicates that average work-in-process increases as average cycle time increases under the same throughput rate. Therefore, how to estimate the cycle time in different system configurations is of great significance to inventory control, due date quoting, and other decision-making in manufacturing systems.

In the literature, estimates of the mean and the variance of cycle time have been extensively studied from many perspectives, including queuing system analysis and simulation output analysis. Bolch *et al.* (1998) summarized the analyses and results under a variety of queuing system settings, such as M/M/1 and GI/M/1 queues. Whitt (1989) analyzed both mean and asymptotic variance of several relevant performance measures in queuing systems and used them to determine simulation runs needed for precise estimation. On the other hand, general statistical simulation output analysis methods can also be applied to estimate the mean and variance of cycle time in the system. For details of these analyses, please refer to Asmussen and Glynn (2007) and the references therein.

However, in many situations, in isolation the mean and the variance of cycle time cannot provide us with an insightful picture of its distribution. It was noted that the distribution of the cycle time can be highly skewed and have heavy tails (McNeil *et al.*, 2005). Therefore, inference based only on the mean and the variance is difficult, especially when the form of the underlying distribution is unknown. For example, a normal distribution with a mean of one and a variance of one is completely different from an exponential distribution with a rate of one, despite them having the same mean and variance. Instead, estimates of different quantiles of the distribution can provide us with a more comprehensive view of the underlying distribution. Several papers in the literature discuss the estimation of the quantiles of cycle time in manufacturing systems. For example, Jain and Chlamtac (1985) developed an algorithm to estimate the specified quantile without individual observation storages. McNeil *et al.* (2005) employed the Cornish–Fisher

*Corresponding author

expansion to estimate cycle time quantiles and also used max/min transformation to improve the accuracy when the distribution significantly deviates from normal distributions. Chen and Kelton (2006) proposed a method that does not rely on prior information about the quantile to be estimated at the cost of large sample sizes.

Despite the importance of these studies, there is one common drawback: they only provide a snapshot of the system performance. In other words, the estimated cycle time quantiles are only valid under the specific configuration of the throughput rate, product mix, buffer size, etc., at which the cycle time observations were collected. However, manufacturing environments are dynamic, which makes the utilization of these estimates difficult. This drawback undermines the role of simulation analysis for decision-making support, where different scenarios need to be considered and compared, as applied in online due date quoting, scheduling, etc. Because of this disadvantage, it is necessary to quantify the relationship between cycle time distributions with other system parameters to make the inference more reliable under changing configurations.

Most of the relevant papers in the literature only consider the mean cycle time as the response variable in a regression on another parameter, such as throughput rates and product mixes (see, for example, Fowler *et al.* (2001), Park *et al.* (2002), and Yang *et al.* (2007)). There is a limited literature on the relationship between cycle time quantiles and throughput rates. Yang *et al.* (2008) assumed that the cycle time followed a generalized gamma distribution and fitted the regression functions of the first three moments of cycle time to the system throughput rates. The parameters of the generalized gamma distribution were determined by making the first three moments equal to the estimated correspondences. Despite the importance of that work, it has certain limitations. First, if the cycle time does not follow a generalized gamma distribution, then the quantile estimate will not be reliably estimated using this methodology. Second, the procedure involves estimation of three cycle time moment–throughput curves and further estimation of gamma distribution parameters based on the previous estimates. Therefore, the variance of the estimated quantile could be significant, and thus more simulation runs are required to obtain the desired accuracy.

In this article, we propose to use quantile regression to directly quantify the relationship between cycle time quantiles and other manufacturing parameters. Particularly, our method offers some advantageous features: (i) it does not rely on distributional assumptions of the cycle time; (ii) it can estimate multiple quantiles simultaneously without additional simulations; and (iii) it can incorporate max/min transformation to reduce the sample size needed for the estimation of extreme quantiles. We will present our method in detail and elaborate these advantages in the following sections.

The remainder of this article is organized as follows. Section 2 gives the formal problem statement while Section 3 presents the methods of estimation and inference of cycle time quantile–throughput curves through quantile regression. Relevant discussions regarding several important issues in practical applications are also provided. Section 4 presents numerical studies to demonstrate the accuracies and efficiencies of the proposed methods. Section 5 concludes the article with discussions on future directions.

2. Problem formulation

To limit the scope of this article, we only consider the regression between cycle time quantiles and system throughput rates. Without loss of generality, we normalize the throughput by the system capacity to be in the range from zero to one and use x_i to denote the ith throughput under consideration. When the system is stationary, the throughput rate is equal to the release rate of new jobs into the system. Therefore, it can be controlled and used as an independent variable in the regression. On the other hand, individual cycle times can also be easily observed by recording the release time and completion time of each job. We use $Y_j(x_i)$ to denote the jth cycle time observation from the simulation where the throughput is set to x_i. It is well known that the cycle time observations from the same simulation are usually dependent, and their distributions may not be the same, especially in the transient period of the system. Therefore, our focus here is on the cycle time quantiles in the long-run behavior during steady state operation of the system. We make the following assumptions on the system under study.

Assumption 1: The process $\{Y_j(x_i)\}$ is stationary for every x_i, and $Y_j(x_i)$ converges in distribution to a random variable $Y(x_i)$ as $j \to \infty$.

Assumption 2: The process $\{Y_j(x_i)\}$ is ϕ-mixing for every x_i. In other words, let $M_{-\infty}^{n}$ and $M_{n+m}^{-\infty}$ be σ-fields generated by $\{Y_j(x_i);\ j \le n\}$ and $\{Y_j(x_i);\ j \ge n+m\}$, if $E_1 \in M_{-\infty}^{n}$ and $E_2 \in M_{n+m}^{-\infty}$, then:

$$|P(E_2|E_1) - P(E_2)| \le \phi(m), \quad \forall n. \tag{1}$$

where $P(\cdot)$ is the corresponding probability measure, and $\phi(m)$ is strictly non-increasing for $m \ge 1$, and $\lim_{m\to\infty} \phi(m) = 0$ (Heidelberger and Lewis, 1984).

These two assumptions are commonly applied in queueing system analysis. Assumption 1 is also made in Yang *et al.* (2008). However, we do not assume the existence of higher moments of the cycle time distributions, which is necessary in Yang *et al.* (2008). It is worth mentioning that although it is difficult for Assumption 1 to be formally proved in general situations, it is fairly intuitive and not restrictive to many applications. Assumption 2 imposes the dependence structure on the output observations. The intuition behind Equation (1) is that the dependence between two samples decreases as the samples become farther apart from each other along the time axis. In another words, samples observed recently can provide more information on

the future observations compared with samples observed earlier. Therefore, if we take samples sufficiently far apart from each other in the steady state as our observations, they can be considered to be independent and identically distributed (i.i.d.) according to the assumptions. It is worth pointing out that regenerative processes such as the waiting time of the M/G/1 system are ϕ-mixing (Heidelberger and Lewis, 1984).

In the literature, the relationship between cycle times and throughputs is often characterized by a non-linear regression function:

$$Y_j(x) = f(x) + \varepsilon_j(x), \tag{2}$$

This model has heteroscedastic errors dependent on x. As the throughput rate x increases, both the mean and the variance of the cycle time will increase. Yang *et al.* (2007) proposed to use a general function inspired from queuing approximation theory and heavy-traffic analysis to model the mean and the variance of the cycle time without batch processing policies. This function takes the form of

$$E[Y_j(x)] = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p} \quad \text{and} \quad var[Y_j(x)] = \frac{\sigma^2}{(1-x)^{2q}}, \tag{3}$$

where m, p, q, β_l, $l = 1, 2, \ldots, m$ are coefficients to be estimated, and σ^2 is the variance of the error term. Similar or simplified functions have been adopted by among others Cheng and Kleijnen (1999) and Park *et al.* (2002). The power function of $(1-x)$ in the denominator captures the fast increasing trend of cycle time as the throughput rate increase, and the polynomial function in the numerator makes the model flexible and able to fit a wide range of distributions. In this article, we conjecture that the quantiles of the cycle time distribution can also be described using the same form. Specifically, we propose to use the Location-Scale Shift (LSS) model to describe the relationship between the cycle time and throughputs:

$$Y_j(x) = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p} + \frac{1}{(1-x)^q}\varepsilon_j, \tag{4}$$

where ε_j are identically distributed with marginal distribution function $F_\varepsilon(t)$. Here, for mathematical simplicity, we make the following further assumption.

Assumption 3: $F_\varepsilon(t)$ is a continuous distribution function with density $f_\varepsilon(t)$, and $0 < f_\varepsilon(t) < \infty$ in the quantile range of interest.

This assumption ensures that the distribution function is invertible, and the quantile function can be simply expressed as the inverse function of the distribution functions. Notably, it is not restrictive, since in most manufacturing environments, cycle time is continuously distributed, which makes Assumption 3 valid. Comparing with other models (e.g., the Expectation Cycle Time model used in Yang *et al.* (2007)), in the LSS model: (i) the error process ε_j can be dependent: we only require it to be ϕ-mixing through Assumption 2; (ii) the marginal distribution of ε_j can be quite arbitrary and is not necessarily normal. These differences make our model extendable to a wider class of production systems.

By definition, the τth quantile of the distribution $F_\varepsilon(t)$, denoted by $Q_\varepsilon(\tau)$, can be expressed as $Q_\varepsilon(\tau) = F^{-1}(\tau) = \inf\{t; F_\varepsilon(t) \geq \tau\}$. By Equation (4) and the equivariance property of non-decreasing monotone transformations of quantiles (Koenker, 2005), we have that the τth quantile for the cycle time under throughput x is

$$Q_{Y(x)}(\tau) = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p} + \frac{1}{(1-x)^q} Q_\varepsilon(\tau). \tag{5}$$

Please note that the model in Equation (5) is over-parameterized with $p, q, m, \beta_l, l = 0, 1, \ldots, m$, and $Q_\varepsilon(\tau)$ as unknown coefficients. In other words, the solution is not unique. As illustrated in the following equation, for an arbitrary value $Q_\varepsilon(\tau)$ and corresponding $p, q, m, \beta_l, l = 0, 1, \ldots, m$, we can always find another set of $p', m', \beta'_l, l = 0, 1, \ldots, m'$ to represent the same function of $Q_{Y(x)}$ with $Q_\varepsilon(\tau) = 0$:

$$\begin{aligned} Q_{Y(x)}(\tau) &= \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p} + \frac{1}{(1-x)^q} Q_\varepsilon(\tau) \\ &= \begin{cases} \dfrac{\left(\sum_{l=0}^{m} \beta_l x^l\right) \cdot (1-x)^{q-p} + Q_\varepsilon(\tau)}{(1-x)^q} \\ \quad = \dfrac{\sum_{l=0}^{m'} \beta'_l x^l}{(1-x)^{p'}} + 0, \quad p \leq q, \\ \dfrac{\sum_{l=0}^{m} \beta_l x^l + (1-x)^{p-q} \times Q_\varepsilon(\tau)}{(1-x)^p} \\ \quad = \dfrac{\sum_{l=0}^{m'} \beta'_l x^l}{(1-x)^{p'}} + 0, \quad p > q. \end{cases} \end{aligned}$$

For example, when $p \leq q$, we have $p' = q$, $m' = m + q - p$, and

$$\beta'_0 = \beta_0 + Q_\varepsilon(\tau), \quad \beta'_l = \sum_{k=0}^{l} \beta_k \times C_{q-p}^{l-k} \times (-1)^{l-k}, \quad l = 1, 2, \ldots, m',$$

where C_n^k is the number of k-combinations from a set of n elements. Thus, to obtain a unique solution, we need to select a value for $Q_\varepsilon(\tau)$ first. Without loss of generality and for the sake of simplicity, we select $Q_\varepsilon(\tau)$ as zero in this paper. In fact, similar treatments have been routinely used in regular linear regressions, where the expectation of the noise term is often assumed to be zero to avoid over-parameterization. With this treatment, our quantile model becomes

$$Q_{Y(x)}(\tau) = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p}, \tag{6}$$

and our objective is: given observations $(x_i, Y_j(x_i))$, $i = 1, 2, \ldots, n$ (n is the number of distinct throughput rates

under consideration), $j = 1, 2, \ldots, N(x_i)$, determine the regression quantile coefficients in location and scale part respectively; i.e., $p, q, m, \beta_l, l = 0, 1, \ldots, m$. Note that the parameter m represents the maximum polynomial order and needs to be specified before any model fitting method is used. In practice, we can specify a large m and then use the model selection technique to select the proper order m.

3. Quantile regression for cycle time modeling

3.1. *Estimation of model parameters*

In this section, we discuss the estimation of the model parameters. We consider the estimation of $p, \beta_l, l = 0, 1, \ldots, m$ given m first in Equation (6) and then the estimation of m and q in Equation (4). Parallel to the least-squares estimation of conditional mean regressions, quantile regression also tries to minimize the distance between observed values and projected values. However, the distance measure in this case becomes the weighted absolute distance, instead of the Euclidean distance in least square estimation. Formally, given m, the parameters in Equation (6) can be estimated by

$$\min_{p,\beta_l} \sum_{i=1}^{n} \sum_{j=1}^{N(x_i)} \left[Y_j(x_i) - \frac{\sum_{l=0}^{m} \beta_l x_i^l}{(1-x_i)^p} \right] \times \left[\tau - I\left(\frac{\sum_{l=0}^{m} \beta_l x_i^l}{(1-x_i)^p} \geq Y_j(x_i) \right) \right], \tag{7}$$

where $I(u) = 0$, otherwise 1 when the statement u is true.

From Equation (7), we can find that the objective function is the sum of the weighed absolute distances in the sense that whenever the observation $Y_j(x_i)$ is smaller than $Q_{Y(x_i)}(\tau)$, the weight on their absolute difference is $1 - \tau$; on the other hand, if $Y_j(x_i)$ is larger than or equal to $Q_{Y(x_i)}(\tau)$, the weight on their absolute difference is τ. In the case without explanatory variables, solving Equation (7) corresponds to using the order statistic to estimate the quantile for identically distributed data. (For readers who are not familiar with quantile regression, a short introduction is included in Appendix A for reference).

Despite the philosophical similarity, there are differences compared to least-squares estimation. In least-squares estimation, the objective function usually has circular contours, and it is differentiable with respect to model parameters if the regression function is differentiable. However, the objective function in Equation (7) only has polyhedral contours, and it is non-differentiable with respect to model parameters due to the discontinuity of the indicator function $I(u)$. Although we cannot obtain an elegant closed-form optimal solution to Equation (7) as in least-squares estimation, a computational efficient estimation method has been proposed to solve the optimization problems similar to Equation (7) (Koenker and Bassett, 1978). We can reformulate Equation (7) to another non-linear programming problem that can be solved using standard approaches, such as the interior-point method:

$$\begin{aligned} &\min \tau \sum_{i=1}^{n} \sum_{j=1}^{N(x_i)} u_{ij} + (1-\tau) \sum_{i=1}^{n} \sum_{j=1}^{N(x_i)} v_{ij}, \\ &\text{subject to } Q_{Y(x_i)}(\tau) + u_{ij} - v_{ij} = Y_j(x_i), \\ &i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, N(x_i), \\ &u_{ij} \geq 0, \quad v_{ij} \geq 0, \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, N(x_i) \end{aligned} \tag{8}$$

The optimization is conducted with respect to u_{ij}, v_{ij}, and model parameters in $Q_{Y(x)}(\tau)$. It can be shown that the optimal solution of $p, \beta_l, l = 0, 1, \ldots, m$ in Equation (8) is also the optimal solution in Equation (7). For details, please refer to Koenker (2005) and the references therein.

To determine the maximum polynomial order m, we can assume a sufficiently large upper bound of m is known, say M, to ensure the goodness of fit of the quantile models. We propose to use stepwise backward elimination with an Akaike Information Criterion (AIC) as the fitness criterion, defined by

$$\mathrm{AIC}(m) = \ln(RS_m) + m + 2, \tag{9}$$

where RS_m is the minimal function value in Equation (7). In Equation (9), the first part $\ln(RS_m)$ indicates how adequate the model can fit the data, while $m + 2$ is the number of unknown parameters estimated. By minimizing the AIC value for different m, we can achieve a balance between goodness of fit and complexity of the model (Akaike, 1974). Different orders of the polynomial functions usually have different approximation accuracies. Also, in many cases, a higher-order polynomial is not necessarily better than a lower-order polynomial with a finite sample size. Therefore, we want to identify the optimal polynomial order that can provide us with a good approximation. To select the best model, we will sequentially compare the AIC value for models with polynomial order k (initially set to M) and $k - 1$. If the higher-order model has a larger AIC value, then we continue comparing the performance of the model with order $k - 1$ and $k - 2$, etc. This procedure continues until further elimination of higher-order terms cannot improve the AIC value.

Using the above-mentioned procedure, we can fully specify the quantile model in Equation (6). However, in many cases, we are also interested in estimating the parameter q in Equation (4). For example, in prediction we are not only interested in the predicted value but also its variance or confidence interval, which relies on the heteroscedastic form of the errors. We can utilize the relationship between the variance of cycle times and the variance of errors to get efficient estimation of q. Specifically, from Equation (4) we

have that:

$$\text{var}[Y(x_i)] = \frac{1}{(1-x)^{2q}} \times \text{var}(\varepsilon) \equiv \frac{\sigma^2}{(1-x)^{2q}},$$

where σ^2 is the variance of ε. A linear regression is conducted on the logarithm of the sample variance based on the variance relationship as

$$\ln S^2(x_i) = \ln \sigma^2 - 2q \ln(1-x_i) + \zeta_i, \tag{10}$$

where $S^2(x_i)$ is the sample variance of $Y_j(x_i)$:

$$S^2(x_i) = \frac{1}{N(x_i)-1} \sum_{j=1}^{N(x_i)} [Y_j(x_i) - \bar{Y}(x_i)]^2 \quad \text{where}$$

$$\bar{Y}(x_i) = \frac{1}{N(x_i)} \sum_{j=1}^{N(x_i)} Y_j(x_i). \tag{11}$$

In Equation (10), x_i is known, and $S^2(x_i)$ can be estimated through Equation (11) for each i; therefore, the unknown parameters $\sigma^2$and q can be easily estimated using linear regression methods. In contrast to some existing approches in which the heteroscedastic error function needs to be estimated before the mean cycle time function can be estimated, our estimation on quantile models does not rely on information about the heteroscedastic error function. Actually, we can even estimate the error function after the quantile model is estimated. The advantage is obvious: if the error model is mis-specified or not estimated correctly, the side effect will not propagate to corrupt the entire estimation of cycle time quantile models.

This overall procedure can be repeated to estimate regression models at different quantile points. However, in general the monotone property of the estimated quantile cannot be guaranteed. In the application where only single or a few quantiles are of interests, this problem is tolerable. If the comparisons among the quantiles under the same setting are also of interest, an alternative procedure to ensure the monotonicity is presented in Appendix A.

3.2. *Confidence intervals on the parameter estimates*

It is of great importance to investigate the properties of the parameter estimates, such as consistency and asymptotic variance. They can be utilized to construct confidence intervals on the true parameters we want to estimate and assess the accuracies of the estimation. Since methods used to estimate m and q are well developed, and their properties are well studied, we concentrate our efforts on the estimation properties of other parameters. Assuming the model structure is correct, and denoting $\Theta = [\beta_0, \beta_1, \ldots, \beta_m, p]^T$ to be the vector of model parameters to be estimated with dimension $m+2$, Θ_0 and $\hat{\Theta}$ to be the true values and corresponding estimates of the parameters obtained from solving Equation (7), respectively, then we have the following theorem.

Theorem 1. *Under Assumptions 1 to 3, and denoting* $N = \sum_{i=1}^{n} N(x_i)$, *if:*

$$\lim_{N\to\infty} \frac{1}{N} \mathbf{W}^{\mathrm{T}} \mathbf{\Phi} \mathbf{W} = \mathbf{K}, \text{ and } \lim_{N\to\infty} \frac{1}{N} \mathbf{W}^{\mathrm{T}} \mathbf{Z} \mathbf{W} = \mathbf{\Omega} \tag{12}$$

where $\mathbf{K}$ *is non-singular,* $\mathbf{W} = [\mathbf{D}_1^{\mathrm{T}}, \mathbf{D}_2^{\mathrm{T}}, \ldots, \mathbf{D}_n^{\mathrm{T}}]^{\mathrm{T}}$, *and* $\mathbf{D}_i$ *is the gradient vector at throughput* x_i:

$$\mathbf{D}_i = \frac{\partial Q_{Y(x_i)}(\tau)}{\partial \Theta_0} = \left[\frac{1}{(1-x_i)^p}, \frac{x_i}{(1-x_i)^p}, \ldots, \frac{x_i^m}{(1-x_i)^p}, \frac{\sum_{l=0}^{m} \beta_l x_i^l}{(1-x_i)^p} \ln\left(\frac{1}{1-x_i}\right) \right]_{\Theta=\Theta_0}$$

$\mathbf{\Phi}$ *and* $\mathbf{Z}$ *are diagonal matrices with the* i*th diagonal entry* $N(x_i) \times f_{Y(x_i)}[Q_{Y(x_i)}(\tau)]$ *and* $\sum_{s=1}^{N(x_i)} \sum_{t=1}^{N(x_i)} [F_\varepsilon^{s,t}(0,0) - \tau^2]$, *respectively.* $F_\varepsilon^{s,t}(0,0)$ *is defined as the joint probability* $\mathrm{P}(\varepsilon_s \le 0, \varepsilon_t \le 0)$ *of the error process from the same simulation output and* τ *is the quantile of interest. Then we have* $\sqrt{N}(\hat{\Theta} - \Theta_0)$ *converges in distribution to a mean* $\mathbf{0}$ *and covariance* $\mathbf{K}^{-1}\mathbf{\Omega}\mathbf{K}^{-1}$ *normal distribution.*

This theorem can be proved by linearizing the model (6) around its true parameter:

$$Q_{Y(x_i)}(\tau|\hat{\Theta}) - Q_{Y(x_i)}(\tau|\Theta_0) = \mathbf{D}_i^{\mathrm{T}}(\hat{\Theta} - \Theta_0) + \varsigma_i, \tag{13}$$

where ς_i is the linearization error, and utilizing the results provided by Oberhofer and Haupt (2005). We omit the details of the proof here and provide some insights of this theorem. In Theorem 1, $\mathbf{\Phi}$ quantifies the impact of heteroscedastic errors on the estimation, where $f_{Y(x)}[Q_{Y(x)}(\tau)]$ is the density of the cycle time at the τth quantile. In the LSS model, we have that:

$$f_{Y(x)}[Q_{Y(x)}(\tau)] = (1-x)^q f_\varepsilon[Q_\varepsilon(\tau)] = (1-x)^q f_\varepsilon(0), \tag{14}$$

by the property of random variable transformation defined through Equation (4). On the other hand, $\mathbf{Z}$ characterizes the influence of dependent error on the estimation. It can be observed that $F_\varepsilon^{s,t}(0,0) - \tau^2 = P(\varepsilon_s \le 0, \varepsilon_t \le 0) - P(\varepsilon_s \le 0) \bullet P(\varepsilon_t \le 0)$, which measures the dependence between two samples. Also, for notation simplicity, we define $F_\varepsilon^{s,s}(0,0)$ to be τ by convention.

In Theorem 1, except for the error density $f_\varepsilon(0)$ and the joint error probability $F_\varepsilon^{s,t}(0,0)$, all the other quantities in $\mathbf{\Omega}$ and $\mathbf{K}$ are either known or can be estimated using the procedure in Section 3.1. Therefore, to construct the confidence interval, we only need to estimate $f_\varepsilon(0)$ and $F_\varepsilon^{s,t}(0,0)$ from data to compute the covariance matrix of the parameter estimations. In terms of estimation of error density $f_\varepsilon(0)$, some work has been performed, e.g., Hendricks and Koenker (1991) and Powell (1991) for i.i.d data, and Ahmad (1979), Hart *et al.* (1990), and Hall *et al.* (1995) for dependent data. In this article, we adopt the kernel estimation method to estimate $f_\varepsilon(0)$ and concentrate our efforts on the estimation of $F_\varepsilon^{s,t}(0,0)$. The following theorem provides us with a way to efficiently estimate the matrix $\mathbf{Z}$.

Theorem 2. *Under Assumptions 1 to 3, the ith diagonal element of matrix* $\mathbf{Z}$ *can be computed as*

$$\sum_{s=1}^{N(x_i)}\sum_{t=1}^{N(x_i)}\left[F_\varepsilon^{s,t}(0,0)-\tau^2\right] = (\tau-\tau^2)\times\left\{N(x_i)+2\sum_{k=1}^{N(x_i)-1}[N(x_i)-k]\times\gamma(k)\right\}, \quad (15)$$

where $\gamma(k)$ *is the autocorrelation function of the process* $I(\varepsilon_j \le 0)$.

The proof of Theorem 2 is included in Appendix B. From Theorem 2, we can find that the estimation of $\mathbf{Z}$ boils down to the estimation of the autocorrelation function $\gamma(k)$. One of the most satisfactory estimates of the kth lag autocorrelation of a process $\{\delta_j\}$ is (Box *et al.*, 1994):

$$\hat{\gamma}(k)=\frac{\sum_{j=1}^{N(x_i)-k}(\delta_j-\bar{\delta})\times(\delta_{j+k}-\bar{\delta})}{\sum_{j=1}^{N(x_i)}(\delta_j-\bar{\delta})^2} \quad \text{where}$$

$$\bar{\delta}=\frac{1}{N(x_i)}\sum_{j=1}^{N(x_i)}\delta_j. \quad (16)$$

We can substitute these auto-covariance estimates to get the estimates of the matrix $\mathbf{Z}$, which represents the effect of dependence on the covariance of parameter estimates.

With all the necessary elements available in $\mathbf{K}$ and $\mathbf{\Omega}$, we can readily compute the variance of the parameter estimates and furthermore evaluate the variance of the predicted quantiles. Through the asymptotic normal property of the model parameters, we can furthermore make inference on the corresponding quantile estimates by using the delta method. Specifically, we have that the estimated quantile values converge in distribution to a normal distribution as:

$$\sqrt{N}[\hat{Q}_{Y(x_i)}(\tau)-Q_{Y(x_i)}(\tau)]\xrightarrow{D}\text{Normal}(\mathbf{0},\mathbf{D}_i^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{\Omega}\mathrm{K}^{-1}\mathbf{D}_i). \quad (17)$$

Therefore, we can construct the confidence interval on the quantile estimates based on the estimation of the covariance matrix. The confidence interval with confidence level $1-\alpha$ is thus:

$$\hat{Q}_{Y(x_i)}(\tau)-z_{\alpha/2}\frac{\mathbf{D}_i^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{\Omega}\mathrm{K}^{-1}\mathbf{D}_i}{\sqrt{N}} \le Q_{Y(x_i)}(\tau) \le \hat{Q}_{Y(x_i)}(\tau) + z_{1-\alpha/2}\frac{\mathbf{D}_i^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{\Omega}\mathbf{K}^{-1}\mathbf{D}_i}{\sqrt{N}}, \quad (18)$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantile points of the standard normal distribution. These confidence intervals are asymptotically valid and can provide us with a more comprehensive view on the accuracy of parameter estimations compared to single point estimates.

3.3. *Practical issues and discussions*

3.3.1. *Cycle time quantiles of batch processing*

Batch processing is a common practice in current manufacturing environments with product varieties, since it can save setup time and increase machine utilization. However, in batch processing, the cycle time does not monotonically increase as the throughput rate increases. Under very low throughputs, objects will spend more time waiting for other objects with which to form a batch; while under very high throughput levels, objects will spend more time waiting in the queue. Therefore, the relationship between cycle time quantiles and throughput rates is typically a U-shaped curve. Park *et al.* (2002) proposed a model for mean cycle time curves with batch processing of the form of

$$Y(x)=\frac{\beta_3}{x}+\frac{\beta_1 x}{\beta_2-x}+\beta_4. \quad (19)$$

Thus, we can also generalize our model (4) to include batch processing production as

$$Y_j(x)=\frac{\sum_{l=0}^{m}\beta_l x^l}{x^{p_1}(1-x)^{p_2}}+\frac{1}{x^{q_1}(1-x)^{q_2}}\varepsilon_j, \quad (20)$$

Obviously, model (20) can characterize the cycle time in cases with or without batch processing. When $p_1=q_1=0$, Equation (20) degenerates to the model (4). However, if prior information about the process is available, it is better not to use the general model as in Equation (20), which could include unnecessary parameters, and hence introduce excessive variance in parameter estimation. If no information is available about the cycle time behavior in the given throughput range, it may be a good strategy to use Equation (20) for preliminary studies of cycle time quantiles.

3.3.2. *Extreme quantiles estimation*

From Theorem 1, we can find that the asymptotic covariance of the estimations is proportional to the inverse of the square error density at the quantile of interest. Mathematically, we have that:

$$\text{var}(\hat{\theta})\propto\frac{1}{f_\varepsilon^2[Q_\varepsilon(\tau)]}. \quad (21)$$

Therefore, when τ is close to zero or one, where the density $f_\varepsilon[Q_\varepsilon(\tau)]$ is small, as in many distributions, the variance of the parameter estimates could be very large. Although we can use the same set of cycle time observations to estimate quantile curves at different quantiles of interests in our methodology, the parameter estimations are expected to have a large variance in extreme quantiles, such as the 95th or 99th quantiles.

Heidelberger and Lewis (1984) and McNeill *et al.* (2005) considered using the max transformation to estimate extreme quantiles for dependent observations obtained from the same simulation. Instead of estimating quantile τ of sequence $\{Y_j\}$, we can estimate quantile τ^ν of sequence $\{T_k\}$, which is generated by $T_k=\{\max Y_{(k-1)\bullet\nu+1}, Y_{(k-1)\cdot\nu+2},$

$\ldots, Y_{kv}\}$. If observation Y_j is i.i.d, then we have $Q_Y(\tau) = Q_T(\tau^v)$. If the observations in the sequence are not independent but are ϕ-mixing, the equality will also approximately hold by ensuring that observations are far from each other in the max transformation, such as $T_k = \max\{Y_{(k-1)\cdot vu+u}, Y_{(k-1)\cdot vu+2u}, \ldots, Y_{kvu}\}$, where u is chosen so that Y_l and Y_{l+u} are approximately independent for $l \geq 1$.

The benefits of the max transformation lie in two aspects. First, if v is chosen to make τ^v close to the median (by setting v to the closest integer of $\ln(0.5)/\ln(\tau)$, then the density of the transformed sequence is usually much larger; i.e., $f_T[Q_T(\tau^v)] > f_Y[Q_Y(\tau)]$. This can reduce the variance of parameter estimates according to Equation (21). Second, the transformation can effectively reduce the dependence among observations, which can also reduce the estimation variance. We will elaborate this point in Section 3.3.3.

In the context of quantile regressions, we can also employ max transformation to reduce estimation variance. First, we can transform all the observations $Y_j(x_i)$ under different throughput to new sequences $T_k(x_i)$ according to the specified quantile τ^v, usually 0.5. Then quantile regressions can be applied on $T_k(x_i)$ as usual, and $Q_{T(x)}(\tau^v)$ would be the estimate of the τth quantile of $Y(x)$, the original cycle times. In the next section, we present a simulation experiment to illustrate the benefits of max transformation in extreme quantile estimation.

3.3.3. *Impact of data dependence on estimation*

One advantage of our method is the ability to handle dependent observations under mild assumptions. However, as in many statistical inference problems, dependent data can usually decrease the estimation efficiency compared with that of independent data with the same sample size. In this section, we would like to investigate how the sample dependence can influence the parameter estimation. From Theorem 2, we know that the ith diagonal element of $\mathbf{Z}$ is actually:

$$\mathbf{Z}_i = (\tau - \tau^2) \times \left\{ N(x_i) + 2 \sum\nolimits_{k=1}^{N(x_i)-1} [N(x_i) - k] \times \gamma(k) \right\}. \tag{22}$$

Therefore, as the auto-covariance function increases, $\mathbf{Z}_i$ increases as well. Furthermore, it can be shown that if $\mathbf{Z}^1 \geq \mathbf{Z}^2$, i.e., $\mathbf{Z}_i^1 \geq \mathbf{Z}_i^2$, $i = 1, 2, \ldots, n$, then the difference between two covariance matrices:

$$(\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}\mathrm{W})^{-1}[\mathbf{W}^{\mathrm{T}}(\mathbf{Z}^1 - \mathbf{Z}^2)\mathbf{W}](\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}\mathrm{W})^{-1}, \tag{23}$$

is a positive semi-definite if $\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}\mathbf{W}$ is invertible. Therefore, the quantile estimates variance under the same throughput will be no smaller in the first case: $\mathrm{var}[Q^1(\tau)] \geq \mathrm{var}[Q^2(\tau)]$. Hence, it is sufficient to investigate the behavior of $\gamma(k)$ to study the variance of quantile estimations from dependent sequences. Obviously, in independent cases, we have $\gamma(k) = 0$ for all $k > 0$. However, if we take the observations of cycle times sequentially from the simulation, they are always correlated. Recall that in GI/G/1 system, the cycle time between two consecutive objects can be expressed as

$$Y_j = \max\{0, Y_{j-1} - A_j\} + P_j, \tag{24}$$

where A_j and P_j are the interarrival time and processing time of the jth object and are independent of Y_{j-1}. It can be shown that Y_j and Y_{j-1} are positively correlated. Furthermore, Pakes (1971) reported that the serial correlation coefficients of the cycle time $\{Y_j\}$ decrease monotonically to zero. Therefore, according to Equation (22), the estimation variance obtained from down-sampled observations with the same sample size is smaller than that without down-sampling. Additionally, it is possible to approximately quantify the effects of correlation on $\mathbf{Z}$. Under stationary condition, denoting the probability $\mathrm{P}(Y_{j-1} \leq A_j)$ by π, we have that:

$$\begin{aligned}
\rho &= \frac{\mathrm{cov}(Y_j, Y_{j-1})}{\mathrm{var}(Y_j)} \\
&= \frac{E(Y) \times E(P) + (1-\pi) \times [E(Y^2) - E(Y) \times E(A)] - (EY)^2}{\mathrm{var}(Y)} \\
&= 1 - \pi + \frac{E(Y) \times [E(P) - (1-\pi) \times E(A) - \pi \times E(Y)]}{\mathrm{var}(Y)},
\end{aligned} \tag{25}$$

where we drop the subscripts j since in a stationary condition, their marginal mean and variance are the same. If we use a first-order autoregressive process AR(1) to approximate the cycle time observations, we have that:

$$Y_j = \rho Y_{j-1} + \zeta_j. \tag{26}$$

Then the autocorrelation function can also be approximated by $\gamma(k) = \rho^k$, which decays exponentially as k increases (Box *et al.*, 1994). Under this approximation, we have that:

$$\begin{aligned}
\mathbf{Z}_i &= (\tau - \tau^2) \times \left\{ N(x_i) + 2 \sum_{k=1}^{N(x_i)-1} [N(x_i) - k] \times \rho^k \right\} \\
&= (\tau - \tau^2) \times \left\{ N(x_i) + 2 \left[\frac{N(x_i) \times \rho}{1-\rho} - \frac{1 - \rho^{N(x_i)}}{(1-\rho)^2} \right] \right\} \\
&\approx (\tau - \tau^2) \times \left[N(x_i) \frac{1+\rho}{1-\rho} - \frac{2}{(1-\rho)^2} \right].
\end{aligned} \tag{27}$$

The last approximation in Equation (27) is due to the fact that $\rho^{N(x)}$ decays to zero for large $N(x_i)$. We can use the results in Equation (27) to quantify the impact of the dependent data. For high traffic intensity queues, it also provides a quantitative guideline on the choice of down-sampling rate. For example, we can solve the inequality:

$$(\tau - \tau^2) \times \left[N(x_i) \frac{1+\rho^\pi}{1-\rho^\pi} - \frac{2}{(1-\rho^\pi)^2} \right] \leq C, \tag{28}$$

with respect to the down-sampling rate π for a given upper bound C on elements in matrix $\mathbf{Z}$.

From the presented analysis, we can find that data dependence indeed plays an important role in parameter estimation. Therefore, in planning simulation experiments, we also need to take this into consideration: shall we allocate the budget on longer simulation and get less dependent observations or on shorter simulation with multiple repetitions. To limit the scope of this article, we do not discuss this point further here.

4. Numerical study

In this section, extensive numerical studies are conducted to demonstrate the effectiveness of the proposed method. In simple systems, theoretical results can be derived for the cycle time quantiles. We compare our estimation with the true value to illustrate the accuracy of our regression model. Also, we use more complicated examples to show that our method can be extended to general systems with satisfactory performances.

4.1. *M/M/1 system with first-come first-serve queues*

4.1.1. *Estimation of quantile curves*

First, we consider the M/M/1 queuing system with a First-Come First-Serve (FCFS) dispatching rule for waiting lines. It is well known from queuing analysis that in steady-state operation the cycle time follows an exponential distribution. Assume that the processing time is exponential distributed with rate λ, and the inter-arrival time is exponentially distributed with rate μ, then under steady-state conditions the system number is distributed as

$$P(L = n) = \left(1 - \frac{\mu}{\lambda}\right)\left(\frac{\mu}{\lambda}\right)^n, \tag{29}$$

where L is the number of objects in the system. By conditioning on the number of existing objects in the system, we can obtain the cycle time distribution as

$$\begin{aligned} P(CT > t) &= \sum_{i=0}^{\infty} P(L = i) P\left(\sum_{j=0}^{i} X_j > t\right) \\ &= \sum_{i=0}^{\infty} \left(1 - \frac{\mu}{\lambda}\right)\left(\frac{\mu}{\lambda}\right)^i \sum_{j=0}^{i} \frac{(\lambda t)^j}{j!} e^{-\lambda t} \\ &= \sum_{j=0}^{\infty} \sum_{i=j}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t} \left(1 - \frac{\mu}{\lambda}\right)\left(\frac{\mu}{\lambda}\right)^i = e^{-(\lambda-\mu)t}. \end{aligned} \tag{30}$$

From Equation (30), we can find that the cycle time is exponentially distributed with rate $(\lambda - \mu)$. Therefore, the τth quantile of the cycle time can be expressed as

$$Q_Y(\tau) = \frac{\ln(1-\tau)}{-(\lambda-\mu)} = \frac{\ln(1-\tau)}{-\lambda(1-x)}. \tag{31}$$

In our simulations, we fixed $\lambda = 1$ and changed the arrival rate (which equals the throughput rate in steady state operation), to get the cycle time observations for model estimation. After a warmup period in each simulation, the output observations were down-sampled by a factor of 10 to reduce the correlation among samples. For illustration purposes, the throughput rates were selected arbitrarily without experimental design consideration. We ran the simulation with ten throughput rates ranging between 0.5 and 0.95, incremented in steps of 0.05. Ten thousand cycle time observations (after down-sampling) under each throughput rate were collected for quantile regression. The simulation and modeling procedure was executed 100 times, and the relative error and absolute error plots are given in Fig. 1. We would like to point out that the computational efforts are mainly spent on the data collection (i.e., simulate the queueing process). With available data, the estimation itself is very fast (computation time is negligible).

From Fig. 1, we can see that the relative errors for all three quantile estimations are almost the same at a given throughput rate. Also, as the throughput increases, the errors increase significantly. Although it seems that the quantile prediction has too much variance, we would like to point out that the sample size here at each throughput is only 10 000. As a matter of fact, McNeill *et al.* (2005) used 1000 000 samples at one throughput to estimate the first four moments of the cycle time distributions for quantile estimation. If we choose the same sample size, then by Theorem 1, the estimation variance will only be one tenth of those in Fig. 1, which are quite small.

Based on the computed variance of the quantile estimates obtained in the process of model estimation, we constructed confidence intervals on the true cycle time quantiles under different throughputs as described in Equation (18). If the estimations are accurate, then the confidence interval should cover the true quantile value with probability $1 - \alpha$. Figure 2 illustrates the coverage frequencies of the constructed confidence interval at confidence level $1 - \alpha = 0.9$ for the quantile $\tau = 0.5$. From Fig. 2, we can observe that the constructed confidence interval has a coverage probability that is slightly higher than the designed value of 0.9. Since the confidence interval is based on the asymptotic covariance estimation for model parameters, for a finite sample size, there are expected to be some discrepancies. Additionally, the estimation of the density and sample dependences may not be accurate enough for the current sample size, which can also cause such differences. However, as long as the difference is tolerable, we can still use this confidence interval to assess the accuracies of the estimation.

4.1.2. *Comparison with other methods*

An alternative simple approach to estimate the quantile curve is to fit a non-linear function of the sample quantile $SQ_\tau(x_i)$ as a function of the throughput rate in the same

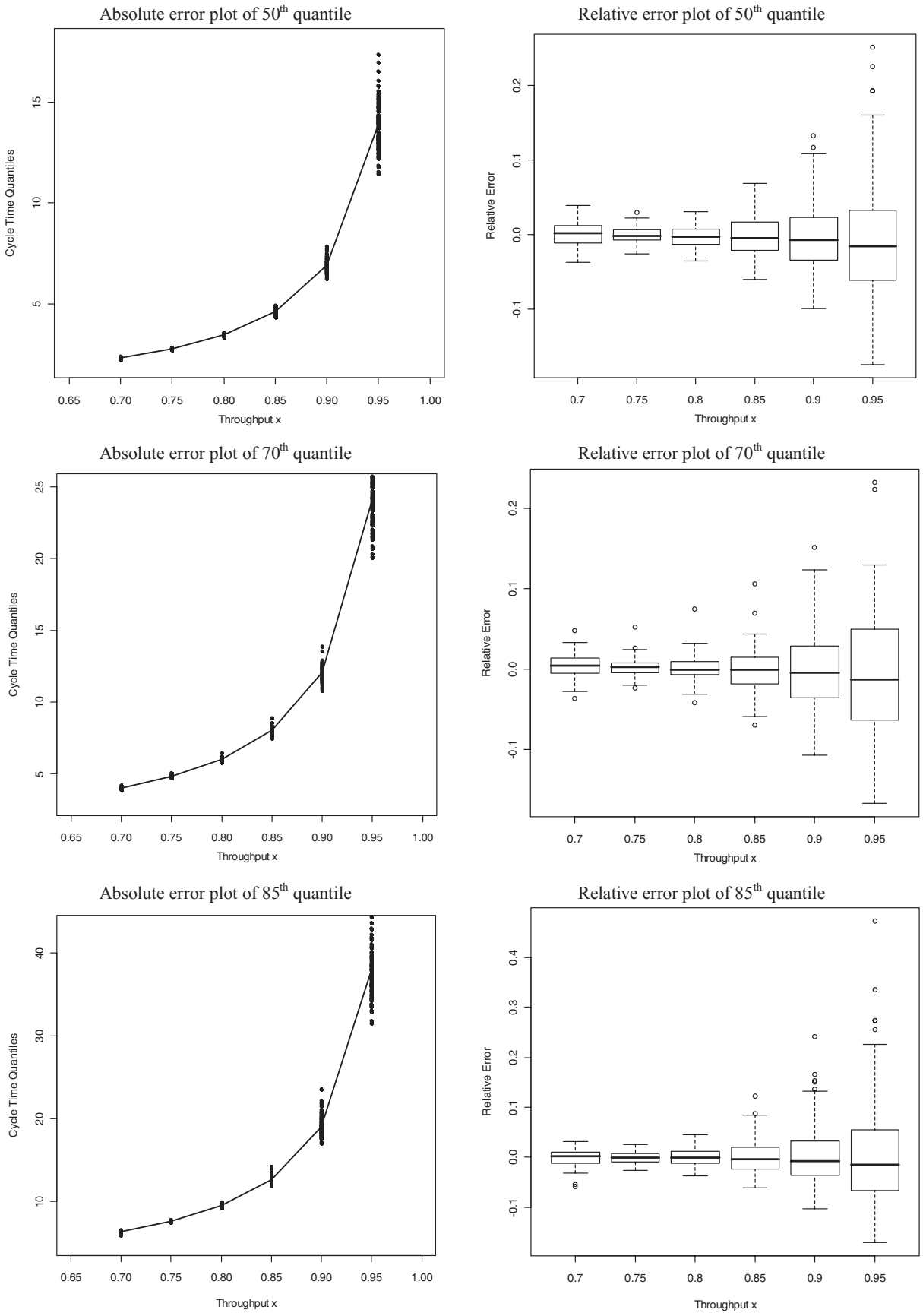

Fig. 1. Quantile estimates for the M/M/1 system.

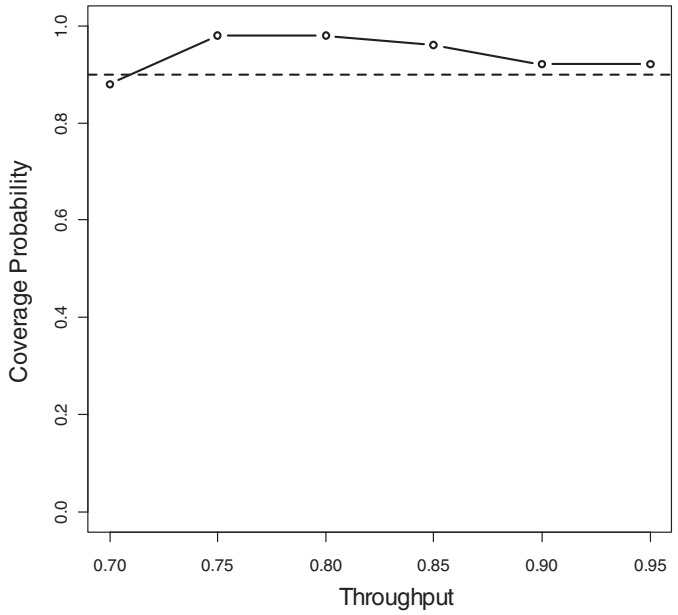

Fig. 2. Coverage probability of constructed confidence interval on 50th quantile estimates.

form as in Equation (6):

$$SQ(x) = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p}. \tag{32}$$

The sample quantile $SQ_\tau(x_i)$ is defined as the corresponding order statistics $Y_{\lfloor \tau \cdot N_i \rfloor}(x_i)$ of the cycle time observations with throughput x_i. We compared the performance of our method with this alternative in terms of prediction error and model selection error.

The left panel in Fig. 3 clearly demonstrates that the prediction error of our proposed method (written as QR) is much smaller than that of the alternative sample quantile fitting method (written as SQ). Additionally, note that the correct order in the numerator of the model for M/M/1 is zero, and QR can consistently select the correct order, whereas SQ has a high probability to select higher orders. This over-fitting phenomenon can partially explain the relatively large prediction error. Furthermore, during our simulation study, it was found that SQ often encounters numerical problems to estimate the parameters, because it only utilizes the summary statistics of all the data in model fitting. Therefore, if we only have ten designed throughput rates in a simulation, we can only get ten samples in estimating possible five or six parameters.

Another method to estimate quantile curves was proposed by Yang *et al.* (2008; shortened to YAN). It essentially estimates the first three moments of the cycle time from the data and computes the quantile curves numerically, with the assumption that cycle times follow a generalized gamma distribution. However, as pointed out in Yang *et al.* (2008, p. 632), "precisely estimating higher moment curves can be very difficult," and "requires substantially more simulation data to obtain well-estimated moment curves." In our simulation study, if we use the same dataset as that in the QR and SQ methods, YAN often finds it difficult to compute the second or third moment curves, and thus it cannot provide meaningful estimates of

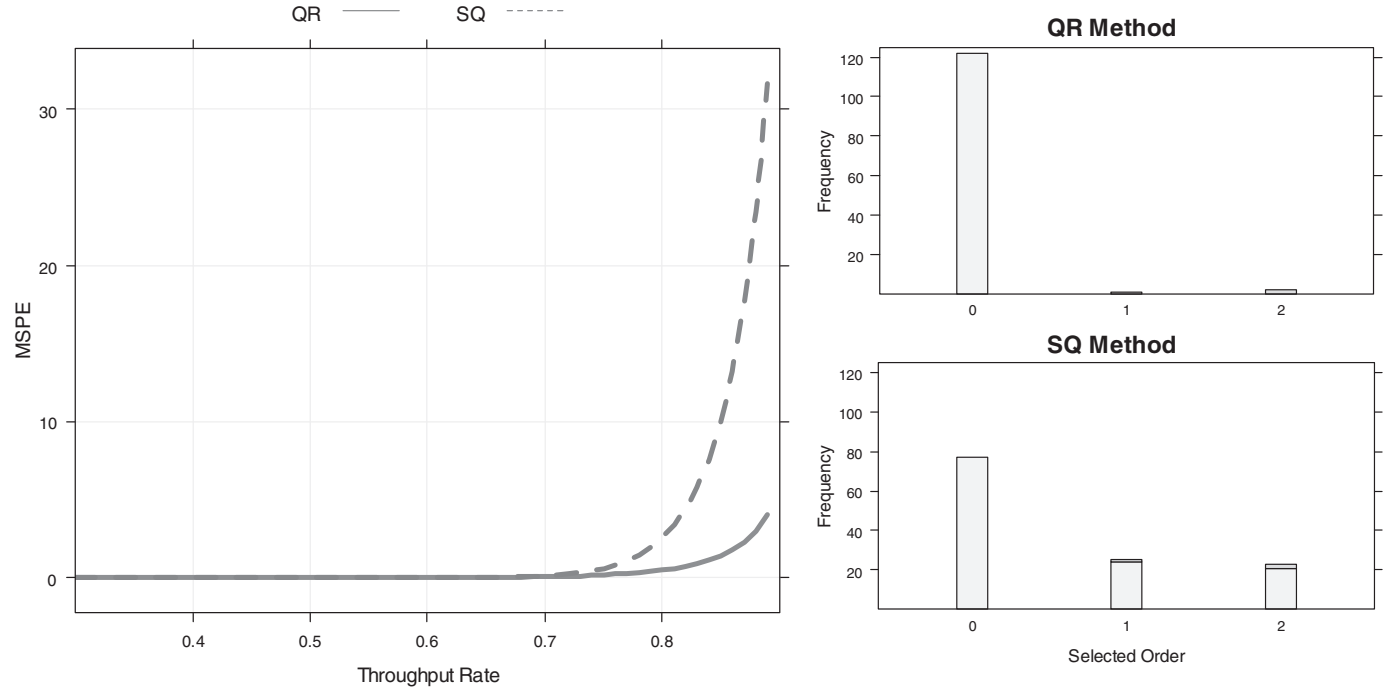

Fig. 3. Comparison of Mean Square Prediction Error (MSPE) and the order of the selected model.

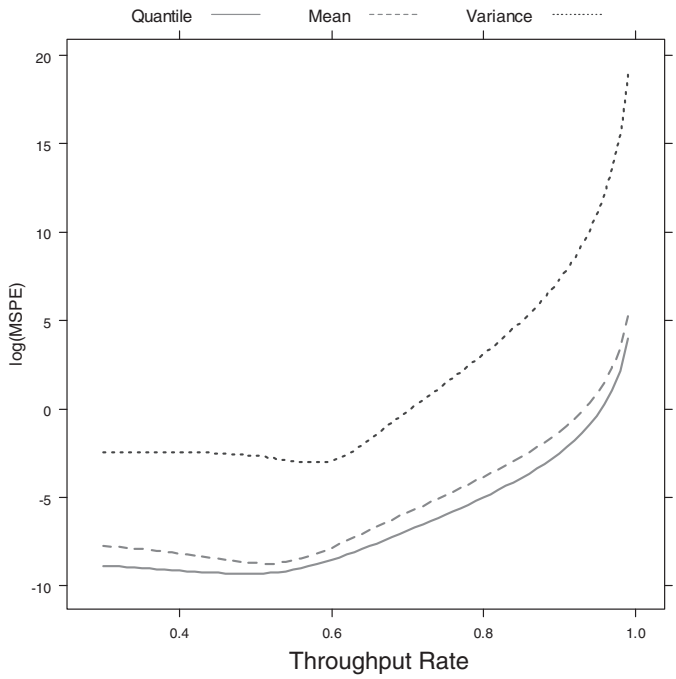

Fig. 4. Comparison of MSPE for quantile, mean, and variance of the cycle time.

the quantile curves. Actually, the sample size necessary for QR to achieve a comparable accuracy is similar (slightly smaller) to that for mean curves, but usually it is significant smaller than that for second and third moments' curves, as demonstrated in Fig. 4. In the figure, the models for quantile, mean, and variance are estimated using the same dataset, and their MSPEs (log scale) are compared. Clearly, the variance estimation has much larger prediction errors than the quantile estimation or mean estimation. It is expected that the third moment would have even larger errors. Therefore, the overall procedure in YAN based on the first three moments would require a very large sample size to ensure estimation accuracy.

4.1.3. *Max transformation for extreme quantiles*

In this experiment, we wanted to estimate the 99th quantile curve within the [0.75, 0.95] range of throughputs of an M/M/1 system. In the first scenario, direct estimation was used on ten design points with 50 000 samples each. In the second scenario, the cycle time observations were collected under the same throughput settings as that in the first scenario, and max transformation was applied to these observations, with ν chosen to be 68 and the corresponding quantile to be 0.505 of the transformed sequence. Table 1 summarizes the results on the Root Mean Square Errors (RMSEs) of the quantile estimates of cycle times under different throughputs. $Q_{Y(x)}(0.99)$ is the theoretical 99th quantile for a cycle time with throughput x; RMSE(Q_{D}) is the root mean square error of quantile estimates under the first scenario; RMSE(Q_{Max}) is the root mean square error of quantile estimates under the second scenario. All the mean square errors were computed from 100 macro runs of the simulations, and we take the square root of them in the table to make the unit consistent with $Q_{Y(x)}(0.99)$. From the comparisons, we can find that with max transformation, the mean square error has been reduced consistently over all the throughput ranges. Notably, with max transformation, we only need 1/68 of the storage space and also the required computational time for quantile regression is much smaller due to the reduced sample size. Therefore, with max transformation, the extreme quantiles can be more efficiently estimated, and its estimation variance can be more effectively reduced.

Table 2. Comparison between estimation variance under different dependences

	0.5	*0.6*	*0.7*	*0.8*	*0.9*	*0.95*
$Q_{Y(x)}$ (0.5)	1.39	1.73	2.31	3.47	6.93	13.86
RMSE (Q_1)	0.19	0.18	0.16	0.15	1.16	6.09
RMSE (Q_2)	0.05	0.05	0.03	0.05	0.33	1.15

4.1.4. *Impact of data dependence on estimation accuracies*

In Section 3.3.3, we analyzed the impact of dependent data on the parameter estimations theoretically. In this section, we would like to demonstrate our analysis through numerical simulation. Table 2 compares the estimation variance obtained with or without down-sampling for the median quantile in a M/M/1 system with FCFS queues. In both cases, the same ten throughput rates were chosen, and 10 000 samples were collected in each simulation. In the first case, the original cycle time observations were used; in the second case, down-sampling was applied to the original observations, and only $Y_{10\cdot j}$, $j = 1, 2, \ldots, 10\ 000$ were taken for parameter estimation.

In Table 2, $Q_{Y(x)}(0.5)$ is the theoretical median cycle time under throughput x, RMSE(Q_1) and RMSE(Q_2) are the RMSEs of the estimated quantiles in the first and second cases, respectively, calculated from 100 repetitions (macro

Table 1. Standard deviation of estimates with/without max transformation

	0.75	*0.77*	*0.79*	*0.81*	*0.83*	*0.85*	*0.87*	*0.89*
$Q_{Y(x)}$ (0.99)	18.42	20.02	21.93	24.24	27.09	30.70	35.42	41.87
RMSE (Q_{D})	3.59	3.78	4.02	4.33	4.73	5.30	6.11	7.37
RMSE (Q_{Max})	2.88	3.04	3.25	3.53	3.91	4.46	5.27	6.55

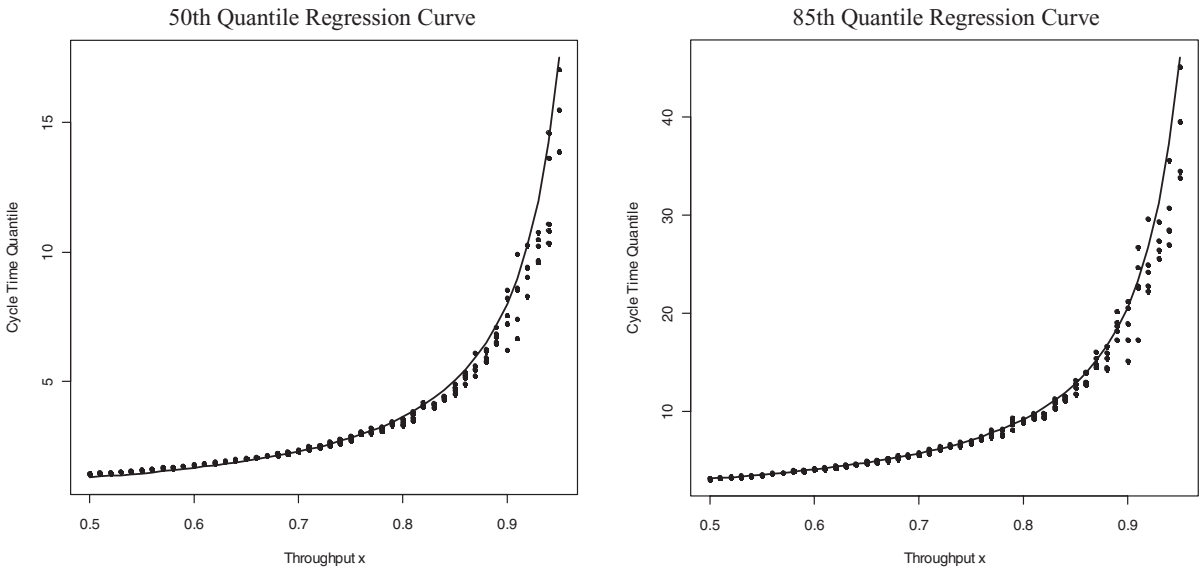

Fig. 5. G/G/1 quantile regression curve with empirical quantile estimates.

runs). From the numbers provided, we can see that the RMSE obtained using down-sampling is only about 20–30% of that from the original observations.

From both analytical and experimental results, it can be observed that the correlations among successive cycle times are much stronger in higher throughput ranges. Therefore, it is possible to devise adaptive down-sampling based on Equation (28) in Section 3.3.3, whose sampling rate is determined by the correlations. In this way, the simulation length needed can be reduced without much sacrifice on the estimation accuracies.

4.2. *G/G/1 system with FCFS queues*

Generally, the G/G/1 queuing model provides more flexibility in approximating real systems compared to M/M/1 queues. However, often the stationary distribution of the cycle time cannot be analytically derived. Therefore, instead of computing the relative error and absolute error between our fitted model and analytical results, we can instead illustrate the prediction accuracy of the regression quantile model. In this experiment, the inter-arrival time was assumed to have a lognormal distribution with the log-variance one and log-mean adjusted according to the throughput requirement. The server processing time was assumed to follow an Erlang (2) distribution with the rate one-half. Therefore, the mean processing time was one in order to be consistent with previous assumptions. As in previous experiments, ten throughput rates equally spaced between 0.5 and 0.95 were selected for the simulation. Ten thousand cycle time observations under each throughput rate were collected for model fitting. A new set of throughputs was chosen ranging from 0.5 to 0.95, incremented in steps of 0.01. Under each throughput, new simulations were conducted and 50 000 observations were collected. The empirical sample quantile was estimated by using the $\lfloor T\tau \rfloor$th order statistic $Y_{\lfloor T\tau \rfloor}$, where T is the sample size (50 000 in this case). At each throughput point, this procedure was repeated five times and the estimated sample quantiles are plotted along with the fitted quantile curves in Fig. 5.

From Fig. 5, we can see that the quantile regression curve can satisfactorily predict the quantiles under different throughput rates. Only in the high throughput range do the predictions have a large variance and are thus not reliable. However, this issue can be solved by using additional replications in the simulations to collect more data for model fitting and thus control the accuracy level of predictions.

4.3. *Serial production lines*

In this section, we consider a serial production system consisting of four workstations. The processing times at each workstation and the inter-arrival times are random variables following general distributions. Buffers exist between two adjacent workstations. The production line is illustrated in Fig. 6.

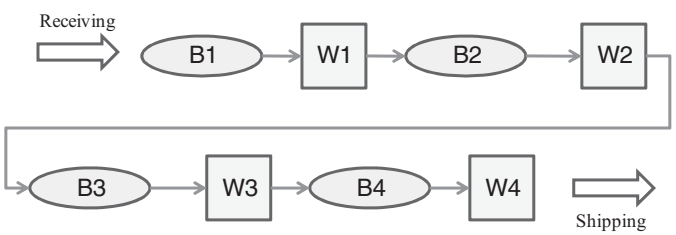

Fig. 6. Illustration of the serial production line.

Table 3. Summary of distribution parameters used in the simulation

	Type	*Mean*	*Var*	*CV*	*Skew*	*Min*	*Max*
W1	Exponential	0.7	0.49	1	2	0	NA
W2	Triangluar	1	0.0417	0.204	0	0.5	1.5
W3	Normal	1	0.04	0.2	0	NA	NA
W4	Uniform	1	0.0833	0.289	0	0.5	1.5
Arrival	Exponential	$1/x$	$1/x^2$	1	2	0	NA

In this figure, "W" represents a workstation, and "B" represents a buffer. It is assumed that the transfer time between workstations and buffers is negligible. The processing time distributions and arrival time distribution assumed in the simulation are summarized in Table 3.

In Table 3, "*CV*" denotes the coefficient of variance, which is defined by the quotient between the standard deviation and the mean of the distribution; "*Skew*" is the third central moment divided by the cubed standard deviation; "*Min*" and "*Max*" are the minimum and maximum values from the distribution and are denoted by "NA" if it is infinity.

In this case study, we considered two scenarios. In the first scenario, the buffer size was assumed to be infinity, and therefore each workstation behaved independently during stationary operations. In the second scenario, we considered finite buffer sizes, and thus the waiting times before each workstation were affected by downstream workstations.

4.3.1. *Cycle time with infinite buffer size*

In the simulations, we considered the throughputs where the mean times between arrivals were 1.5, 1.4, 1.3, 1.2, 1.15, 1.1, and 1.05. The simulation was run for 100 000 time units and the first 20% of observations were discarded as the warm-up period. Different to the simulations in Section 4.1 and 4.2, a different number of observations for different throughputs were used in this case.

Using the method described in Section 3, we modeled the relationship between the quantiles of the cycle time and throughputs. Here, we chose the quantiles of interests to be $\tau = 0.5$ and $\tau = 0.85$. To demonstrate the effectiveness of the modeling, sample quantiles were also computed directly using the observed sequences under different throughputs (possibly different from the design points). The estimated models as well as sample quantiles are compared in Fig. 7. From Fig. 7, we can see that the proposed model can closely approximate the true cycle time quantiles, although it tends to overestimate it for high throughput rates. The

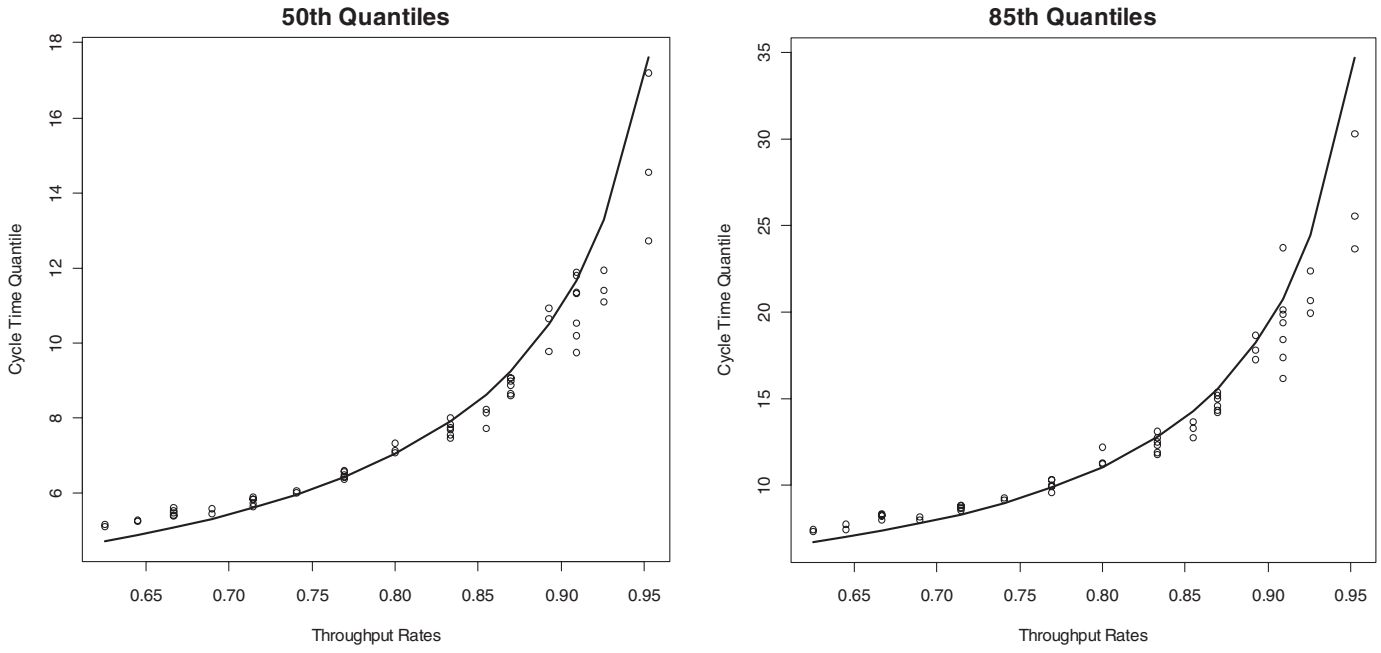

Fig. 7. Estimated quantiles curves of cycle time with infinite buffers.

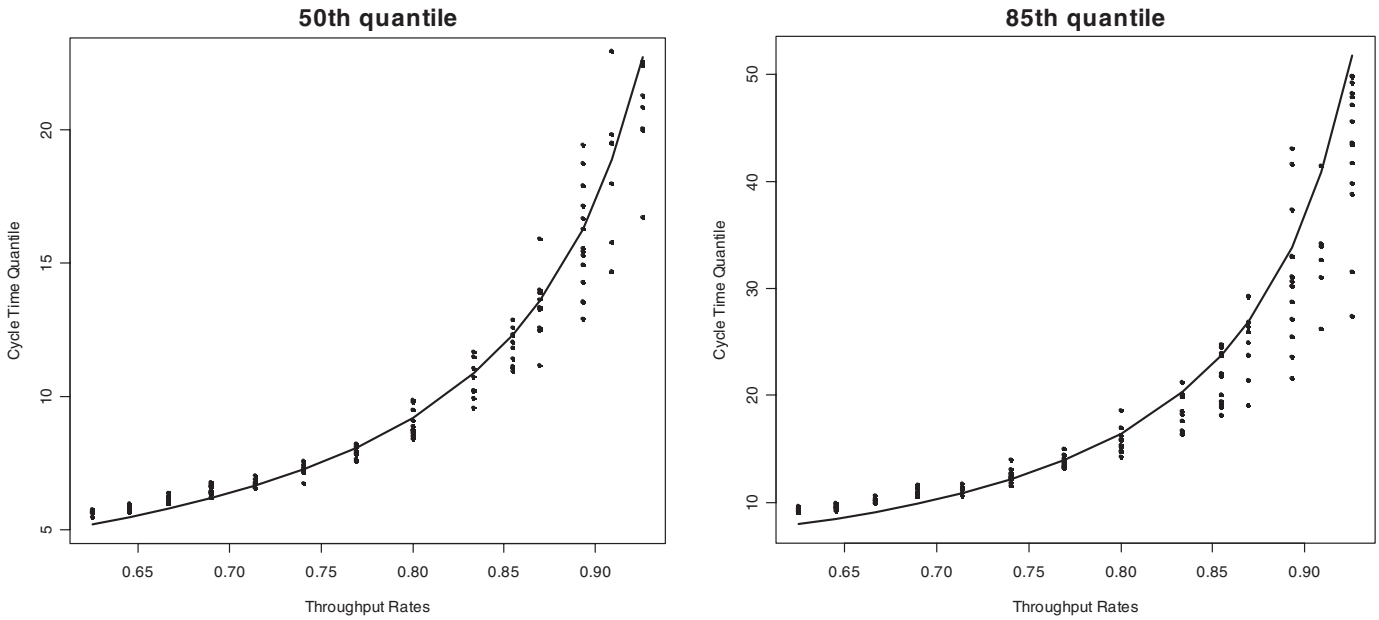

Fig. 8. Estimated quantiles curves of cycle time with ten buffers.

variance of the estimation can be reduced by taking more simulation runs or increasing the run length in each simulation. As previously mentioned, in infinite buffer cases, the workstations act as being independent, and thus the results should be pretty similar to the results in G/G/1, as presented in Section 4.2.

4.3.2. *Cycle time with finite buffer size*

In the second scenario, we considered finite buffer sizes between workstations, which is a more realistic representation of real industrial environments. The buffer size before W1 was assumed to be infinite, which means that no input orders could be rejected, although some of them may wait for a long time. The buffer size between successive workstations was set to ten. The buffer size in the shipping area was again taken to be infinite. We used the same set of throughput rates as in the simulations in Section 4.3.1. It was expected that the cycle time would be longer than that in the infinite buffer cases, since the downstream buffers can block the upstream workstations from processing, thereby increasing the total waiting time in the queue. This phenomenon is especially obvious in the system with high throughputs.

Figure 8 summarizes the 50th and 85th quantile curve estimates under different throughputs. From Fig. 8 it can be observed that the predicted curves fit the sample quantiles quite well despite the large variance in the high throughput range. Compared to the cycle time quantile curves in Fig. 7, we find that the cycle time quantiles increases dramatically under high throughputs with finite buffers. On the other hand, under relatively small throughputs, the cycle time quantiles are almost the same in the two scenarios. These observations are consistent with our intuitions described in the previous paragraph.

From the experiments in the case studies, we can find that the proposed cycle time quantile model and quantile regression methods can be successfully applied to a wide class of production systems with satisfactory accuracy and efficiency. We can thus utilize these estimated models to support a variety of decision-making activities in manufacturing environments.

5. Conclusions and future work

In this article, we propose a quantile regression model to characterize the relationship between cycle time quantiles and system throughput rates. The properties of estimations and inferences are also studied theoretically and experimentally. Numerical case studies have been presented to illustrate the effectiveness and accuracy of the proposed method, which also demonstrate that our methods are broadly applicable to many production environments. We also provided short discussions on some practical issues that may occur in applying our methods.

There are still some open issues in the proposed technique. For example, how to assess and control the estimation precisions efficiently is worth more investigations. In Yang *et al.* (2008), adaptive design and simulation was introduced to allocate the simulation budget and was shown to be effective. In the future, we will also study the optimal experimental designs in the context of quantile regression. In other words, with limited budget, we want to have a systematic procedure to choose the optimal design points and corresponding run length or run numbers to get a good estimation with high accuracy.

Moreover, although we have successfully quantified the relationship between cycle time quantiles and throughput rates, many other important factors in the production environment can also significantly influence the cycle time. Therefore, it is worth extending the current approach to

include other influential factors, such as product mix and order routing paths into the model to more precisely characterize the distribution of cycle times. We will also explore along this direction in the future.

Acknowledgements

The authors appreciate the editors and reviewers for their valuable comments and suggestions. This research is supported by NSF grants #0757683 and #0545600.

References

Ahmad, I.A. (1979) Strong consistency of density estimation by orthogonal series methods for dependent variables with applications. *Annals of the Institute of Statistical Mathematics: Part A*, **31**(1), 279–288.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Asmussen, A. and Glynn, P. (2007) *Stochastic Simulation: Algorithms and Analysis*, Springer-Verlag, New York, NY.

Bolch, G., Greiner, S., Meer, H. and Trivedi, K.S. (1998) *Queuing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, John Wiley & Sons, New York, NY.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994) *Time Series Analysis: Forecasting and Control*, third edition, Prentice Hall, Englewood Cliffs, NJ.

Chen, E.J. and Kelton, W.D. (2006) Quantile and tolerance-interval estimation in simulation. *European Journal of Operational Research*, **168**(2), 520–540.

Cheng, R.C.H. and Kleijnen, J.P.C. (1999) Improved design of queuing simulation experiments with highly heteroscedastic responses. *Operations Research*, **47**(5), 762–777.

Fowler, J.W., Park, S., Mackulak, G.T. and Shunk, D.L. (2001) Efficient cycle time-throughput curve generation using a fixed sample size procedure. *International Journal of Production Research*, **39**(12), 2595–2613.

Gordon, V.S. (1993) A note on optimal assignment of slack due dates in single machine scheduling. *European Journal of Operational Research*, **70**(3), 311–315.

Hall, P., Lahiri, S.N. and Truong, Y.K. (1995) On bandwidth choice for density estimation with dependent data. *The Annals of Statistics*, **23**(6), 2241–2263.

Hart, J.D. and Vieu, P. (1990) Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, **18**(2), 873–890.

Heidelberger, P. and Lewis, P.A.W. (1984) Quantile estimation in dependent sequences. *Operations Research*, **32**(1), 185–209.

Hendricks, W. and Koenker, R. (1991) Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, **87**(417), 58–68.

Hopp, W.J. and Spearman, M.L. (1996) *Factory Physics: Foundations of Manufacturing Management*, Irwin/McGraw-Hill, Chicago, IL.

Jain, R. and Chlamtac, I. (1985) The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Communication of the ACM*, **28**(10), 1076–1085.

Koenker, R. (2005) *Quantile Regression*, Cambridge University Press, New York, NY.

Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**(1), 33–50.

McNeill, J.E., Fowler, J.W., Mackulak, G.T. and Nelson, B.L. (2005) Cycle-time quantile estimation in manufacturing systems employing dispatching rules, in *Proceedings of the 37th Winter Simulation, Conference*, Orlando, FL, pp. 749–755.

Oberhofer, W. and Haupt, H. (2005) Nonlinear quantile regression under dependence and heterogeneity. Regensburger diskussionsbeiträge zur wirtschaftswissenschaft, Department of Economics, University of Regensburg, Regensburg, Germany.

Pakes, A.G. (1971) The serial correlation coefficients of waiting times in the stationary GI/M/1 queue. *The Annals of Mathematical Statistics*, **42**(5), 1727–1734.

Park, S., Fowler, J.W., Mackulak, G.T., Keats, J.B. and Carlyle, W.M. (2002) D-Optimal sequential experiments for generating a simulation-based cycle time-throughput curve. *Operations Research*, **50**(6), 981–990.

Powell, J.L. (1991) Estimation of monotonic regression models under quantile restrictions, in *Nonparametric and Semiparametric Methods in Econometrics*, Barnett, W.A., Powell, J.L. and Tauchen, G.E. (eds), Cambridge University Press, Cambridge, UK, pp. 357–384.

Whitt, W. (1989) Planning queuing simulations. *Management Science*, **35**(11), 1341–1366.

Yang, F., Ankenman, B. and Nelson, B.L. (2007) Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics*, **54**(1), 78–93.

Yang, F., Ankenman, B. and Nelson, B.L. (2008) Estimating cycle time percentile curves for manufacturing systems via simulation. *INFORMS Journal on Computing*, **20**(4), 628–643.

Appendix

A. A Brief Introduction to Quantile Regression

Quantile regression, originally proposed by Koenker and Bassett (1978), has been widely used to measure differences among a family of distributions at multiple points. Similar to least-squares regression, quantile regression also tries to identify the relationship between dependent variables y and covariates $\mathbf{x}$. Consider the following regression model:

$$y = f(\mathbf{x}) + \varepsilon. \tag{A1}$$

It is well known that the optimal function minimizing the loss function $E[y - f(\mathbf{x})]^2$ is the conditional mean function $E[y|\mathbf{x}]$. Correspondingly, it can be proved that the optimal function minimizing the loss function $E\{[y - f(\mathbf{x})] \times [\tau - I(f(\mathbf{x}) \geq y)]\}$ is the conditional τth quantile function $Q_\tau(y|\mathbf{x})$. Therefore, for a finite sample size N we can minimize the empirical loss function:

$$\sum_{i=1}^{N} [y_i - f(\mathbf{x}_i)] \times [\tau - I(f(\mathbf{x}_i) \geq y_i)]. \tag{A2}$$

to obtain a good estimate of the conditional quantile function of interest. This procedure can be repeated for different τ to get regression curves at different quantile points.

In general, multiple estimations at different quantile points cannot guarantee the monotone property of the estimated quantiles. This phenomenon has been studied in the literature (the so-called quantile crossing in Koenker (2005)). Koenker (2005) pointed out that although the monotone property in the full range of x (the regression covariates) may be implausible, the quantile curves within a certain design region are usually monotone. Additionally, violations of the monotone property in the

design region can serve as a good indicator of model inadequacy.

In our particular example, an alternative method can be used to guarantee the monotone property. Note that in our model:

$$Q_{Y(x)}(\tau) = \frac{\sum_{l=0}^{m} \beta_l x^l}{(1-x)^p} + \frac{1}{(1-x)^q} Q_\varepsilon(\tau), \qquad \text{(A3)}$$

when $0 < x < 1$, $q > 0$, and $Q_\varepsilon(\tau)$ is a monotone function of τ, then $Q_{Y(x)}(\tau)$ is ensured to be monotone with respect to τ. Therefore, the estimation for multiple quantiles can work in the following two steps.

Step 1. Select a single quantile τ_0 and use quantile regression to estimate the parameters $p, q, m, \beta_l, l = 0, 1, \ldots, m$. Then the residuals can be computed using:

$$\varepsilon_{ij} = \left[Y_j(x_i) - \frac{\sum_{l=0}^{m} \beta_l x_i^l}{(1-x_i)^p} \right] \times (1-x_i)^q. \qquad \text{(A4)}$$

Step 2. When estimating other quantile points $\tau_1 < \tau_2 < \cdots < \tau_k$, we can compute the sample quantile of the residuals using the order statistics, which ensures the monotone property of the residual's quantiles $Q_\varepsilon(\tau_1) < Q_\varepsilon(\tau_2) < \cdots < Q_\varepsilon(\tau_k)$. Then using Equation (A3) we can get the monotone estimates of $Q_{Y(x)}(\tau_1) < Q_{Y(x)}(\tau_2) < \cdots < Q_{Y(x)}(\tau_k)$.

Although this procedure can produce monotone estimates of multiple quantiles, there is some efficiency loss in this procedure because of the systematic error possibly introduced in Step 1. Detailed discussion and comparisons between the proposed method in Section 3 and the method here are beyond the scope of this article. Interested readers can choose from these two methods based on their application requirements.

B. Proof of Theorem 2. By the definition of $F_\varepsilon^{s,t}(0,0)$ and the stationarity assumption on ε_j, we have that.

$$\begin{aligned} F_\varepsilon^{s,t}(0,0) &= P(\varepsilon_s \le 0, \varepsilon_t \le 0) = E[I(\varepsilon_s \le 0, \varepsilon_t \le 0)] \\ &= E[I(\varepsilon_s \le 0) \times I(\varepsilon_t \le 0)]. \end{aligned} \qquad \text{(A5)}$$

Therefore, we can construct a new binary sequence $\{\delta_j,\ j = 1, 2, \ldots, N(x_i)\}$ from the original error process $\{\varepsilon_j,\ j = 1, 2, \ldots, N(x_i)\}$ by

$$\delta_j = \begin{cases} 1 & \text{if } \varepsilon_j \le 0 \\ 0 & \text{if } \varepsilon_j > 0 \end{cases} \qquad \text{(A6)}$$

By definition, the mean and the variance of δ_j can be computed as

$$E(\delta_j) = P(\varepsilon_j \le 0) = \tau, \quad \text{var}(\delta_j) = E\left(\delta_j^2\right) - [E(\delta_j)]^2 = \tau - \tau^2. \qquad \text{(A7)}$$

Since the sequence $\{\delta_j\}$ is stationary by the stationarity of ε_j, the auto-covariance function $\text{cov}(\delta_s, \delta_t)$ is only a function of $s - t$, and the autocorrelation function $\gamma(k)$ is defined by $\text{cov}(\delta_s, \delta_{s+k})/\text{var}(\delta_s)$. Thus, the expectation in Equation (A5) can be represented by

$$\begin{aligned} E[I(\varepsilon_s \le 0) \times I(\varepsilon_t \le 0)] &= \text{cov}(\delta_s, \delta_t) + E(\delta_s) \\ &\quad \times E(\delta_t) = \text{cov}(\delta_s, \delta_t) + \tau^2. \end{aligned} \qquad \text{(A8)}$$

With $F_\varepsilon^{s,s}(0,0)$ defined as τ by convention, the ith diagonal element of matrix $\mathbf{Z}$ can be represented by

$$\begin{aligned} &\sum_{s=1}^{N(x_i)} \sum_{t=1}^{N(x_i)} [F_{s,t}(0,0) - \tau^2] \\ &= N(x_i) \times (\tau - \tau^2) + \sum_{s=1}^{N(x_i)} \sum_{t=1, s \ne t}^{N(x_i)} \gamma(s-t) \times \text{var}(\delta) \\ &= N(x_i) \times (\tau - \tau^2) + 2 \sum_{k=1}^{N(x_i)-1} [N(x_i) - k] \times \gamma(k) \times (\tau - \tau^2) \\ &= (\tau - \tau^2) \times \left\{ N(x_i) + 2 \sum_{k=1}^{N(x_i)-1} [N(x_i) - k] \times \gamma(k) \right\} \end{aligned} \qquad \text{(A9)}$$

which completes the proof. ■

Biographies

Nan Chen is an Assistant Professor in the Department of Industrial and Systems Engineering at National University of Singapore. He obtained his B.S. degree in Automation from Tsinghua University, and M.S. degree in Computer Science, M.S. degree in Statistics, and Ph.D. degree in Industrial Engineering from University of Wisconsin–Madison. His research interests include statistical modeling and surveillance of service systems, simulation design and modeling, quality control and improvement. He is a member of INFORMS and IIE.

Shiyu Zhou is an Associate Professor in the Department of Industrial and Systems Engineering at the University of Wisconsin–Madison. He received his B.S. and M.S. in Mechanical Engineering from the University of Science and Technology of China in 1993 and 1996, respectively, and his master's in Industrial Engineering and Ph.D. in Mechanical Engineering from the University of Michigan in 2000. His research interests include in-process quality and productivity improvement methodologies by integrating statistics, system and control theory, and engineering knowledge. His research is sponsored by the National Science Foundation, Department of Energy, Department of Commerce, and industries. He is a recipient of a CAREER Award from the National Science Foundation and the Best Application Paper award from *IIE Transactions* in 2006. He is a member of IIE, INFORMS, ASME, and SME.