Operations Research Letters 50 (2022) 529–535

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Point and interval estimation of decomposition error in discrete-time open tandem queues

Christoph Jacobi *, Kai Furmans

Karlsruhe Institute of Technology, Institute for Material Handling and Logistics, Gotthard-Franz-Str. 8, 76131, Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 1 November 2021
Received in revised form 18 July 2022
Accepted 22 July 2022
Available online 29 July 2022

Keywords:
Decomposition
Tandem queue
Waiting time
Multiple linear regression
Quantile regression
ANOVA

ABSTRACT

We analyze the approximation quality of the discrete-time decomposition approach, compared to simulation, and with respect to the expected value and the 95th-percentile of waiting time. For both performance measures, we use OLS regression models to compute point estimates, and quantile regression models to compute interval estimates of decomposition error. The ANOVA reveal major influencing factors on decomposition error while the regression models are demonstrated to provide accurate forecasts and precise confidence intervals for decomposition error.

1. Introduction

Queueing models are widely used for performance evaluation of production and logistics systems which are subject to the influence of randomness [23,33,35,39,41]. When applying continuous-time queueing models, engineers calculate the first and second moment of performance indicators of interest (e.g. throughput, waiting time, and the number of customers in the queue) using the well-known formulas for M/M/1 and M/G/1 queues as well as approximation formulas for G/G/1 queues. Books that provide an overview of continuous-time queueing models are written by Buzacott and Shanthikumar [6] and Wolff [38].

However, production and logistics systems are typically designed to guarantee performance not on average, but with a given probability (e.g. 95%), which necessitates the calculation of the distribution of key performance indicators (such as waiting time) to know, for example, which percentage of orders is processed in 3h or less, or what promised throughput time will be met in 95% of the cases [29]. Applying discrete-time queueing models allows for the computation of the entire probability distributions of key performance indicators under very general assumptions. Discrete-time modelling means that events are only recorded at moments that are multiples of a constant time unit t_{inc}. Thus, the probability mass function of a discrete random variable x is denoted by

$$P(x = i \cdot t_{inc}) = x_i \quad \forall i = 0, \ldots, i_{max}.$$

Given the discrete random variables for the inter-arrival and service time, the probability distributions of performance measures can be computed, for example the waiting time [10] or the inter-departure time [15] distributions. A comprehensive introduction to discrete-time queueing models can be found in [1,5]. The models have been successfully applied in various use cases related to logistics and production systems [7,27–30].

The analysis of discrete-time open queueing networks relies on a decomposition approach. As in the continuous-time domain, the technique is known to yield approximate results in the case of non-Poisson arrivals and generally distributed service times. The drawback with approximations is that we cannot quantify the deviation of the performance measures calculated with a decomposition approach from their actual values. While the approximation quality of decomposition approaches has been studied in the literature for the continuous-time domain (see e.g. [16,34]), decomposition error in the discrete-time domain has not yet been comprehensively examined. So far, no estimator is available to predict decomposition error for a given queueing network in the discrete-time domain.

In this paper, we investigate discrete-time open tandem queues to analyze and forecast the approximation quality of the discrete-time decomposition technique, compared to simulation. We limit ourselves to the analysis of tandem queues with external Poisson arrivals that become non-renewal at the downstream queue with the aim to reveal fundamental dependencies regarding the approximation quality of the discrete-time decomposition approach.

* Corresponding author.
E-mail address: jacobi@kit.edu (C. Jacobi).

https://doi.org/10.1016/j.orl.2022.07.009

2. Theoretical background

Open queueing networks allow for the analysis of systems with infinite buffer capacity and generally distributed inter-arrival and service times. Generalizations of Jackson's product form solution [12,13] with respect to generally distributed inter-arrival and service times are proposed by Reiser and Kobayashi [24] with modifications presented by Kuehn [22], Shantikumar and Buzacott [32], Whitt [36], and Bitran and Tirupati [3,4]. Each decomposition approach relies on two basic assumptions [8]: First, it is assumed that the individual queueing systems can be treated as being statistically independent GI/G/1-queues. Second, it is assumed that the point process which forms the input to each GI/G/1-queue can be approximated by a renewal process. It is therefore important to emphasize that congestion measures obtained by decomposition techniques are approximate, since the assumption of independence among queueing systems does not properly account for the correlations of the arrival stream which have a significant effect on the performance measures [16].

Decomposition approaches for discrete-time open tandem queues are based on these conditions, as well. The arrival stream of a downstream queue is approximated as renewal process by the inter-departure time distribution of the upstream queue, which can be efficiently computed with the algorithm by Jain and Grassmann [15]. The waiting time distribution of the resulting GI/G/1-queue is obtained with the algorithm presented by Grassmann and Jain [10]. Further performance measures, such as the distribution of customers, can be computed with the approaches presented by Hasslinger [11], and Grassmann and Tavakoli [9].

In an effort to investigate the approximation quality of the decomposition techniques, tandem lines have been studied extensively in the literature. Suresh and Whitt [34] examine the impact of non-renewal processes on the approximation quality with different traffic intensities. Wu and McGinnis [40] introduce the intrinsic ratio, a fundamental property of tandem queues that is based on the insight that some servers are directly affected by the external arrival process. Whitt [37] suggests using a variability function (instead of a single parameter as in the QNA) for the arrival stream of the downstream queue, which is a function of the traffic intensity of the incoming queue. Sagron et al. [25] extend this method to multi-class systems that address the scenario when the upstream server in a tandem queue experiences downtimes (e.g. set-up, maintenance, and repair), events that increase the station's departure variability, while causing starvation of a downstream bottleneck station. To achieve better computational efficiency, Sagron et al. [26] approximate the between-class effect (the variability caused by interactions with other classes) in a queue with downtimes using a Regression-Based Variability Function (RBVF). RBVF receives the squared coefficient of variation of the arrival and service times, as well as the expected value of the service process as input and approximates the variability function using methods of linear regression.

3. Methodology

The object of investigation in this paper is a tandem queue, that is, two discrete-time queueing systems are arranged one after the other. The upstream queueing system is fed by an external arrival stream with arrival rate $1/E(A_u)$ of customers. If the service station is busy upon arrival of a customer, this customer waits for service in the waiting room. After being served at the upstream station with service rate $1/E(B_u)$, all customers enter the waiting area of the downstream queueing system. The size of the waiting area is infinite, meaning that all customers wait to be served with service rate $1/E(B_d)$ at the downstream station and to hereafter leave the tandem queue. We only consider steady-state systems where the utilization parameters $\rho_u = E(B_u)/E(A_u)$ and $\rho_d = E(B_d)/E(A_d)$ are smaller than 1. Since the arrival process at the downstream queue is approximated as point process with inter-arrival time distribution A_d, only the downstream queueing system is prone to decomposition error. For the sake of clarity, Table 1 defines the system performance metrics of the tandem queue.

Table 1
Performance metrics of the tandem queue.

A_u, A_d	Random variable describing the inter-arrival time of the external (downstream) arrival process
B_u, B_d	Random variable describing the service time at the upstream (downstream) queue
ρ_u, ρ_d	Utilization of the upstream (downstream) queue
W	Random variable describing the waiting time of a customer at the downstream queue

In our analyses, we assume that the random variables describing the service processes are described by discretized gamma distributions. Let X be a gamma-distributed random variable with shape parameter k and scale parameter θ. The probability density function of X is given by [2]

$$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x,k,\theta > 0,$$

where $\Gamma(k)$ is the gamma function. We use the squared coefficient of variation (scv) as normalized measure of statistical dispersion to measure the process variability. Let $E(X)$ define the expected value of X, and $Var(X)$ its variance. The variability of X is defined as

$$scv(X) = Var(X)/E^2(X).$$

In order to generate gamma-distributed random variables X with predefined values for $E(X)$ and $scv(X)$, we use the well-known closed-form expressions for the shape and scale parameters of the gamma distribution,

$$E(X) = k\theta,$$
$$Var(X) = k\theta^2.$$

Finally, we define σ_X^τ as the τ-percent percentile of the probability mass function (pmf) of random variable X.

In this paper, we are interested in the error of the waiting time W at the downstream queue computed by the discrete-time decomposition approach, compared to discrete-event simulation. We conduct two distinct studies with different dependent variables. In Study I, let $\Delta(E)$ be the divergence of the expected value of waiting time

$$\Delta(E) = \frac{E_{Sim}(W) - E_{Queue}(W)}{E_{Sim}(W)}, \tag{1}$$

where $E_{Sim}(W)$ and $E_{Queue}(W)$ denote the expected value of waiting time, computed with the discrete-time queueing approach and simulation, respectively. In Study II, let $\Delta(\sigma)$ be the divergence of the 95th-percentile of waiting time

$$\Delta(\sigma) = \frac{\sigma^{95}_{W,Sim} - \sigma^{95}_{W,Queue}}{\sigma^{95}_{W,Sim}}, \tag{2}$$

where $\sigma^{95}_{W,Queue}$ denotes the 95th-percentile of waiting time, computed with the discrete-time queueing approach, and $\sigma^{95}_{W,Sim}$ the 95th-percentile of waiting time, obtained with simulation.

In the following, we introduce the methodologies used for the computation of point and interval estimates of decomposition error and briefly describe the empirical evaluation criteria, the simulation model, and our design of experiments.

3.1. Point and interval estimates

We use Ordinary Least Square (OLS) multiple linear regression to compute point estimates, and quantile regression to compute interval estimates for decomposition error. The methodological background on OLS regression can be found e.g. in [31]. Quantile regression aims at the estimation of conditional quantile functions-models in which quantiles (percentiles) of the conditional distribution of the dependent variable are expressed as functions of observed covariates [17,20]. Unlike OLS which is used to compute the conditional mean of the dependent variable, quantile regression can be used to explain the determinants of the dependent variable at any point of the pmf of the dependent variable.

The dependent variables of the regression models in Study I and Study II are $\Delta(E)$ and $\Delta(\sigma)$, respectively. In both studies, we consider the same sample of N observations for the estimation of decomposition error. To help simplify the notations introduced in the following, we do not differentiate between both studies, but instead set $y_n = \Delta(E)$ in Study I, and $y_n = \Delta(\sigma)$ in Study II for a given data point n. The observations include $\boldsymbol{y}$ and $\boldsymbol{X}$, where $\boldsymbol{y}$ denotes the N-vector of decomposition error, and $\boldsymbol{X}$ is the $(N \times K)$ design matrix of the independent variables, with $K-1$ dependent (explanatory) variables.

Point estimates for decomposition error are computed with the well known formula for multiple linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon}$ is the N-vector of the random error terms of the regression model. The estimates $\hat{\boldsymbol{\beta}}$ for problem (3) are found by minimizing the sum of squares residuals

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^K} \sum_{n=1}^{N} \left(y_n - \boldsymbol{x}_n^{\intercal} \boldsymbol{\beta} \right)^2.$$

In contrast to OLS, quantile regression finds the estimates $\hat{\boldsymbol{\beta}}(\tau)$ for a given quantile $\tau \in (0,1)$ by minimizing the weighted sum of the absolute deviations

$$\hat{\boldsymbol{\beta}}(\tau) = \min_{\boldsymbol{\beta}(\tau) \in \mathbb{R}^K} \sum_{n=1}^{N} \left| y_n - \boldsymbol{x}_n^{\intercal} \boldsymbol{\beta}(\tau) \right| \omega_n, \tag{4}$$

where the weight ω_n is defined as

$$\omega_n = \begin{cases} 2\tau & y_n - \boldsymbol{x_n}^{\intercal} \boldsymbol{\beta}(\tau) > 0, \\ 2 - 2\tau & \text{otherwise.} \end{cases}$$

The quantile regression estimates $\hat{\boldsymbol{\beta}}(\tau)$ in problem (4) can be computed very efficiently by linear programming methods. In this paper, we use the modified version of Barrodale and Roberts algorithm [18,19] to calculate the quantile regression estimates.

We always consider the quantile regression models in pairs, so that they form the upper and lower endpoints of the 90%, 95% or 99% confidence interval (CI) of decomposition error, respectively. Consequently, we fit quantile regression models $Q(\tau)$ for the pairs of $\tau = .05$ and $\tau = .95$ for the 90% CI, $\tau = .025$ and $\tau = .975$ for the 95% CI, and $\tau = .005$ and $\tau = .995$ for the 99% CI.

3.2. Goodness of fit criteria and likelihood ratio tests

To evaluate the accuracy of the fitted OLS models, we are interested in the empirical distribution of the error term $\boldsymbol{\varepsilon}$ in problem (3). A preliminary evaluation of the data set shows that the Gauss-Markov conditions [31], and especially $E(\boldsymbol{\varepsilon}) = 0$, hold for our data set. Consequently, mean error measurements for the cumulated error terms of $\boldsymbol{\varepsilon}$ (such as *MSE* and *RMSE*) will be (nearly) zero and therefore not meaningful for interpretation. Instead, we evaluate the absolute values $|\varepsilon_n|, n \in N$ and denote $|\varepsilon_n|$ as *forecasting error (FE)* for observation n. To arrive at the determination of the accuracy of the OLS model, we compute the relative frequency distribution function of *FE* for all observations in $\boldsymbol{\varepsilon}$. Interpreting the relative frequency distribution of *FE*, the higher the percentage of small values, the better the model fits the data and thus the higher the accuracy of the model.

The goodness of fit criterion of quantile regression is calculated with the algorithm by Koenker and Machado [21]. Analogous to the conventional R^2 statistic of OLS regression, we call it Pseudo R^2. Let $\hat{\boldsymbol{\beta}}(\tau)$ denote the minimizer of problem (4), and $\hat{V}(\tau)$ the error sum of the conditional quantile function. Further, let $\tilde{V}(\tau)$ denote the error sum of the corresponding conditional quantile function, that is restricted to only consider the intercept parameter of $\hat{\boldsymbol{\beta}}(\tau)$. Conventionally, the goodness of fit criterion is defined as

$$R^2_{Pseudo}(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau).$$

Note that Pseudo R^2 is not comparable to the standard coefficient of determination R^2 although it lies between 0 and 1. It is only useful for the comparison between quantile regression models since it is based on the weighted sum of absolute residuals, while R^2 is based on residual variance. Finally, it should be noted that Pseudo R^2 may be a skewed measure as it is not corrected by the degrees of freedom. However, a definition for the goodness of fit that follows the concept of Adjusted R^2 known from OLS regression is not available for quantile regression analyses.

We use likelihood ratio tests to test the overall significance of the OLS regression models [31]. We are interested in testing whether all the independent variables have any effect on decomposition error and test the general linear hypothesis

$$H : \boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{\gamma} = \boldsymbol{0}, \tag{5}$$

where $\boldsymbol{C}$ is a $(M \times K)$ matrix of rank $M < K$ and $\boldsymbol{\gamma}$ is a M-vector. Note that hypothesis (5) allows us to test the overall significance of the OLS model, where

$$H : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_{(K-1)} = 0, \tag{6}$$

as well as the significance of elected independent variables (so-called nested models), where

$$H : \beta_m = \gamma_m, \tag{7}$$

for arbitrary values of m and γ_m. Hypothesis (5) is rejected if

$$\frac{M^{-1}(\boldsymbol{C}\boldsymbol{b} - \boldsymbol{\gamma})^{\intercal}[\boldsymbol{C}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{C}^{\intercal}]^{-1}(\boldsymbol{C}\boldsymbol{b} - \boldsymbol{\gamma})}{s^2} \geq F_{M,N-K-1,\alpha}, \tag{8}$$

where $F_{M,N-K-1,\alpha}$ is the upper α-percent point of the F-distribution with $(M, N-K-1)$ degrees of freedom,

$$\boldsymbol{b} = (\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y}, \text{ and}$$
$$s^2 = (N-K-1)^{-1}\boldsymbol{y}^{\intercal}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}]\boldsymbol{y}.$$

We report the test statistic (8) as well as the p-value of the hypothesis test, which is the probability of observing a value of F larger than the one observed under H with degrees of freedom $(M, N-K-1)$ and significance level α. Generally speaking, when the test statistic is large, and the p-value is small, we can safely reject H and conclude that the OLS model provides a better fit to the data than a model which contains no independent variables (hypothesis (6)) or the nested model (hypothesis (7)).

3.3. Simulation model

We use a discrete-event simulation model of a tandem queue to obtain the waiting time distribution at the downstream station. Each simulation run is composed of 50 replications with 10,000,000 simulated time steps each. In each simulation run, the first 100,000 time steps are discarded. The observed width of the 95%-CI of the expected value of waiting time is 0.0286, which is less than 0.5% of the average simulated waiting time. The authors therefore feel that the performance metrics obtained with the simulation model – despite being prone to some variance – are valid estimates for the waiting time.

3.4. Design of experiments

Each tandem queue is parameterized with rate and variability parameters of the external arrival stream and the service processes in both queueing systems. For the sake of conciseness, we limit ourselves to experiments where the arrival process at the first queue is Poisson, and the service times at both queues are gamma-distributed. Given its flexibility, the gamma distribution allows for the modelling of a wide range of dispersion and is therefore well suited to represent the stochastic behaviour of the service process. Further, it is well known that the exponential distribution is a special case of the gamma distribution when the scv-value equals 1. We first consider tandem queues where the utilization parameters at the upstream and the downstream queue are equal. This allows us to define a generic utilization parameter ρ for the tandem queue, $\rho = \rho_u = \rho_d$. A relaxation of this assumption will be discussed in Section 5.

Based on these conditions, we parameterize each tandem queue with four parameters, the external arrival rate, the service rate, and the variability parameters of both service processes. We define the utilization of the tandem queue ρ, the variability parameters of both service processes $scv(B_u)$ and $scv(B_d)$, and the variability of the arrival process at the downstream queueing system $scv(A_d)$ as independent variables (IVs) of the regression models. We partition the data set into two subsets, the *training data set* which consists of 932 randomly chosen data points, and the *test data set* which consist of the remaining 234 data points. The data sets are accessible in a repository [14] and described in detail in the accompanied data article.

4. Results

We first consider the distribution of decomposition error in the overall data set. The empirical cumulative distribution of decomposition error reveals that both, positive (meaning that discrete-time queueing theory underestimates the waiting time) and negative errors (overestimation of the waiting time) are found. We find the relative errors in the range of -21.9% and 32.5% (referring to Study I) and -30.8% and 36.7% (referring to Study II). The mean absolute values of decomposition error equal 3.93% and 4.51% regarding the expected value and the 95th-percentile of waiting time, respectively.

4.1. Study I: Expected value of waiting time

The OLS regression coefficients for Study I are presented in the accompanied data article. Recall that in Study I, the dependent variable is $\Delta(E)$, cf. equation (1). The OLS regression analysis is found to be statistically significant ($F(10, 921) = 2123,\ p < .001$), explaining the majority of the variance of the relative error of the expected value of waiting time ($R^2_{Adj.} = 0.958$). The ANOVA reveals all direct and the majority of the interaction effects to be statistically significant. Since the non-significant coefficient is small, we did not find evidence for the regression model to perform significantly better without incorporating this interaction ($F(921, 922) = 1.234,\ p = .267$). We identify the service process variability at the upstream queueing system and the arrival process variability at the downstream queueing system, as well as the utilization as major impact factors. Despite being statistically significant, the effect of the variability of the service process at the downstream queueing system is found to be a minor influencing factor.

The Pseudo R^2 of each quantile regression model is well above 0.8. All quantile regression equations show similar patterns of changes in coefficient values as the OLS regression. We find the majority of direct and interaction effects to be statistically significant. As in the OLS regression, the interaction effect between the service process variability (at the upstream queueing system) and the utilization is found to be non-significant among each model. While the absolute sizes of the coefficients for most factors vary little across the equations, it should be noted that the weights of the service process variability at the upstream queueing system, and the arrival process variability at the downstream queueing system rise with increasing quantile.

4.2. Study II: 95th-percentile of waiting time

The regression coefficients for Study II are presented in the accompanied data article. In Study II, the dependent variable is $\Delta(\sigma)$, cf. equation (2). We find a statistically significant OLS regression equation ($F(10, 921) = 1064,\ p < .001$), which explains the majority of the variance ($R^2_{Adj.} = 0.920$) of decomposition error regarding the 95th-percentile of waiting time. The impact patterns of the interaction effects are the same as in Study I. Again, we did not find evidence for the OLS estimate to better perform without incorporating the non-significant interaction effect between the service process variability and utilization ($F(921, 922) = 0.917$, $p = .339$). Analogous to Study I, the service process variability (at the upstream queueing system), the arrival process variability (downstream queueing system), and the utilization are found to be the major direct effects. Despite being statistically significant, the service process variability at the downstream queueing system is a minor impact factor.

The Pseudo R^2 of all quantile regression models is well above 0.6. Except for the service process variability at the downstream queueing system, which is non-significant for the models with $\tau \leqslant .05$, all direct effects are found to be statistically significant among each regression model. The majority of interaction coefficients is found to be significant or marginally significant. However, we did find non-significant coefficients among the interaction effect of the service process variability and the arrival process variability (both at the downstream queueing system), as well as in the $Q(.975)$ model. As in Study I, the absolute sizes of coefficients vary little for most factors across the equations. However, the weight of the utilization increases by rising quantiles, while (in contrast to Study I) the weight of the arrival process variability decreases.

4.3. Performance of point and interval estimates

The accuracy of the point estimates is presented in Table 2. For the majority of data points, we find an absolute error of the OLS predictions of less than 1 percentage point from the simulated value. The mean absolute forecasting errors are less than 1 percentage point in Study I and only slightly above 1 percentage point in Study II. In both studies, this accuracy is achieved for the training and the test data set, which indicates that our OLS prediction approach is robust to overfitting.

Despite the minor mean errors, the results suggest that the accuracy of point estimates decreases when forecasting severe values

Table 2
Performance of point estimates: Relative frequency distributions and means of forecasting error for training and test data in Study I and Study II.

FE	Study I				Study II			
	Train	Test	Test (a)	Test (b)	Train	Test	Test (a)	Test (b)
[0.000, 0.005]	40.5%	37.2%	41.9%	8.7%	40.8%	30.8%	34.6%	11.8%
(0.005, 0.010]	30.5%	30.8%	34.6%	13.0%	36.5%	29.5%	33.1%	0.0%
(0.010, 0.020]	20.5%	22.6%	18.4%	34.8%	7.4%	23.9%	19.7%	35.3%
(0.020, 0.050]	8.0%	9.0%	5.1%	39.1%	12.7%	12.8%	11.8%	41.1%
(0.050, ∞)	0.5%	0.4%	0.0%	4.4%	2.6%	3.0%	0.8%	11.8%
Mean	0.0087	0.0092	0.0073	0.0210	0.0118	0.0117	0.0095	0.0260

Notes: Subsets (a) and (b) denote the subsets of test data with absolute decomposition error smaller than 3% and above 10%. The sample sizes are 136 and 23 (Study I), and 127 and 33 (Study II).

Table 3
Performance of interval estimates: Mean lengths and actual share of values for confidence intervals (CI) in Study I and Study II, based on quantile regression models.

	90% CI		95% CI		99% CI	
	Length	Share	Length	Share	Length	Share
Study I Train	0.0348	90.58%	0.0416	95.07%	0.0506	98.93%
Study I Test	0.0345	86.32%	0.0415	91.45%	0.0509	94.02%
Study I Test (a)	0.0271	87.50%	0.0318	93.38%	0.0374	94.85%
Study I Test (b)	0.0692	78.26%	0.0810	82.61%	0.1006	91.30%
Study II Train	0.0467	90.26%	0.0661	95.50%	0.1343	98.82%
Study II Test	0.0473	91.45%	0.0658	92.74%	0.1303	95.73%
Study II Test (a)	0.0432	92.91%	0.0594	96.85%	0.1442	99.21%
Study II Test (b)	0.0689	81.82%	0.1006	81.82%	0.1281	84.85%

Notes: Subsets (a) and (b) denote the subsets of test data with absolute decomposition error smaller than 3% and above 10%. The sample sizes are 136 and 23 (Study I), and 127 and 33 (Study II).

of decomposition error. To investigate this effect, we examine the subsets of test data with minor decomposition errors, that is, all data points with absolute decomposition errors smaller than 3% (in the following referred to as subset (a)), and with severe decomposition errors, that is, all data points with absolute decomposition errors above 10% (subset (b)). The sample sizes of subsets (a) and (b) are 136 and 23 in Study I, and 127 and 33 in Study II, respectively. The relative frequency distributions of *FE* and its mean errors (cf. Table 2) suggest that subset (a) is forcasted with significantly higher accuracy than the data points from subset (b) in both studies. Further, the share of data points that is forecasted with a *FE* greater than 0.05 is significantly higher in subset (b). However, it cannot be concluded that data points with severe absolute decomposition errors are frequently predicted with minor accuracy. In the test data from Study I, we find 96% of the data points with an absolute decomposition greater than 10% to be forecasted with a *FE* less than 0.05 (in Study II the share is 89%).

Interval estimation compensates for this effect. By providing the 90%, 95%, and 99% confidence intervals, we evaluate the precision of the point estimates. Table 3 presents the performance of the interval estimates for Study I and Study II, listing the mean interval lengths and the actual shares of decomposition errors included in the respective confidence intervals. As expected, the average interval lengths increase with rising confidence in finding a data point in the corresponding interval. In both studies, the average interval lengths differ only marginally between training and test data which indicates that the approach of interval estimation is robust to overfitting. In the training data set, the confidence intervals contain exactly the respective share of values they were determined for. These shares are only slightly undermined for the test data.

The interval estimates are designed to indicate uncertainty in the forecast of point estimates. The results are presented in Table 3. In subset (a), the precision of interval estimations increases, compared to the entire test data set. This is indicated by the narrower intervals, as well as the high shares of values that are included in the respective intervals (which is especially to be emphasized for Study II). As discussed above, in subset (b), the forecast uncertainty of the point estimates increases, which is indicated by longer mean intervals and a smaller share of values contained in the intervals.

We conclude that minor decomposition errors are predicted with satisfactory point estimation accuracy and great precision. Predicting severe decomposition errors is subject to uncertainty: the absolute error of the point estimate might be considerable, which is indicated by large confidence intervals. By combining the methods, the authors feel to satisfy both, the aspect of an accurate point estimation forecast, as well as the quantification of its uncertainty.

5. Bottlenecks and longer lines

The investigations of the heavy-traffic bottleneck phenomenon in open queueing systems [34] suggest that the performance of bottleneck downstream queues is strongly related to the variability of the non-renewal arrival process variability, which impacts the approximation quality of decomposition methods. Therefore, we extend our analyses to tandem queues and longer lines with bottlenecks. As Suresh and Witt [34] mention, in a narrower sense, the bottleneck is the queue with the highest traffic intensity. However, increasing the traffic intensity of a queue by only a small amount may shift the bottleneck position. Therefore, it is intuitive to state that either of the queues is the bottleneck if it's utilization is substantially greater than some ϵ, $|\rho_u - \rho_d| > \epsilon$.

We create a further data set containing 969 data points, following the procedure described in section 3.4, but with the relaxation that the expected values of service times are now independent. We choose $\epsilon = 0.1$ and find 403 data points where the downstream queue is the bottleneck. We use OLS and quantile regression to identify the major and minor effects on decomposition error in bottleneck queues. The coefficients of the regression analyses, where the dependent variables are $\Delta(E)$ (Study I), and $\Delta(\sigma)$

Table 4
Absolute mean decomposition errors for Study I and Study II in longer lines.

Queue	Length 3		Length 5		Length 7		Length 9	
	Study I	Study II	Study I	Study II	Study I	Study II	Study I	Study II
1	0.23	0.39	0.24	0.26	0.24	0.22	0.17	0.16
2	2.43	2.43	2.16	2.32	2.43	2.55	2.18	2.39
3	8.94	10.94	2.56	2.38	2.61	2.39	2.72	2.67
4			2.50	2.45	3.12	3.05	2.82	2.83
5			9.68	12.53	3.04	3.22	2.86	3.11
6					3.24	3.36	3.49	3.28
7					9.24	10.67	3.54	4.10
8							3.48	3.03
9							10.91	12.91

Note: Values for decomposition error in percent.

(Study II) are provided in the accompanied data article. We find the previously identified major and minor effects on decomposition error to apply in this analysis, as well. However, the empirical distributions of the decomposition error show that the approximation quality of the decomposition approach depends significantly on which of the queues is the bottleneck. In the case of similar traffic intensities, we find mean absolute values of decomposition error to be 5.45% (6.51%) for the expected value (95th-percentile) of waiting time, which is in line with the expectations of previous examinations. When the bottleneck is downstream, the mean absolute values of decomposition error regarding the expected value (95th-percentile) of waiting time equal 4.87% (5.50%). In contrast, when the bottleneck is upstream, we find mean absolute values of decomposition error of 1.36% (1.46%) for the expected value (95th-percentile) of waiting time.

Similar results are observed in longer lines. We investigate a set of lines with i queues in series, where i equals 3, 5, 7, and 9. For each line length i, we evaluate 250 data points. The utilization parameters of the first $i-1$ queues are equal, and the last queue in each case is the bottleneck. Table 4 shows the mean absolute decomposition errors for the expected value (Study I), and the 95th-percentile (Study II) of waiting time. It can be clearly seen that the last queues are prone to significant decomposition errors with 9.69% on average in Study I, and 11.67% on average in Study II. This is significantly more than the decomposition errors for the intermediate queues which are 2.82% on average in Study I, and 2.85% on average in Study II. The results confirm the long-range variability effect formulated by Suresh and Whitt [34], that states that variability in the external arrival stream or the service times can have a dramatic effect on a downstream queue with a much higher traffic intensity.

6. Concluding remarks

From the analyzes of decomposition techniques in the continuous-time domain, it is well known that utilization and variability parameters for arrival and service processes are significant for the approximation quality of congestion measures. Based on the regression coefficients, we identify utilization and arrival process variability as major impact factors on decomposition error. Service process variability was revealed as a minor impact factor.

Utilization is found to be the enabler for decomposition error: In low-traffic queueing systems, the mean absolute decomposition error is significantly lower than the mean absolute errors in the entire data set. Severe absolute decomposition errors are only observed in heavy-traffic systems. In tandem queues with bottlenecks, we find the decomposition error to be significantly higher when the bottleneck is downstream. This leads to the conclusion that downstream bottlenecks are analyzed with limited accuracy, which should be of particular interest since the performance evaluation of bottlenecks is obviously particularly critical. The arrival process variability determines the tendency (that is, overestimation or underestimation of the waiting time) of the decomposition technique. For scv-values of the arrival process at the downstream queue lower than 1.0, the decomposition approach underestimates waiting time. Overestimation of waiting time occurs for scv-values of the downstream arrival process greater than 1.0. Variability of the service process is a minor impact factor. This is indicated by the fact that when the arrival process at the downstream queue is Poisson, we did not find considerable decomposition errors, regardless of the utilization of the queueing system nor the scv-value of the service process.

We conclude the discrete-time decomposition approach to analyze low traffic queueing systems with high accuracy. In heavy-traffic systems, the approximation quality depends on the arrival process variability. The analysis of queueing systems with highly volatile as well as deterministic arrival processes is prone to considerable decomposition errors. When the arrival process is Poisson, the decomposition approach yields high accuracy, regardless of the service process variability.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

Data are available in a repository (cited in the manuscript).

Acknowledgements

The authors wish to thank Dr.-Ing. Uta Mohring for her many helpful comments on an earlier version of this manuscript.

References

[1] M. Ackroyd, Computing the waiting time distribution for the G/G/1 queue by signal processing methods, IEEE Trans. Commun. 28 (1) (1980) 52–58, https://doi.org/10.1109/TCOM.1980.1094582.
[2] F. Bijma, M. Jonker, A.W.v.d. Vaart, Introduction to Mathematical Statistics, 2017.
[3] G.R. Bitran, D. Tirupati, Multiproduct queueing networks with deterministic routing: decomposition approach and the notion of interference, Manag. Sci. 34 (1) (1988), https://doi.org/10.1287/mnsc.34.1.75.
[4] G.R. Bitran, D. Tirupati, Capacity planning in manufacturing networks with discrete options, Ann. Oper. Res. 17 (1989) 119–135, https://doi.org/10.1007/BF02096601.
[5] H. Bruneel, B.G. Kim, Discrete time models for communication systems including ATM, Kluwer International Series in Engineering and Computer Science, Kluwer, Boston, 1993.
[6] J. Buzacott, J. Shanthikumar, Stochastic Models of Manufacturing Systems, Prentice Hall International Series in Industrial and Systems Engineering, Prentice Hall, 1993.
[7] M. Epp, S. Wiedemann, K. Furmans, A discrete-time queueing network approach to performance evaluation of autonomous vehicle storage and retrieval systems, Int. J. Prod. Res. 55 (4) (2017) 960–978, https://doi.org/10.1080/00207543.2016.1208371.

[8] M.K. Govil, M.C. Fu, Queueing theory in manufacturing: a survey, J. Manuf. Syst. 18 (3) (1999) 214–240, https://doi.org/10.1016/S0278-6125(99)80033-8.
[9] W. Grassmann, J. Tavakoli, The distribution of the line length in a discrete time GI/G/1 queue, Perform. Eval. 131 (2019) 43–53, https://doi.org/10.1016/j.peva.2019.03.001.
[10] W.K. Grassmann, J.L. Jain, Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic GI/G/1 queue, Oper. Res. 37 (1) (1989) 141–150.
[11] G. Haßlinger, A polynomial factorization approach to the discrete time GI/G/1/(N) queue size distribution, Perform. Eval. 23 (3) (1995) 217–240, https://doi.org/10.1016/0166-5316(94)00024-E.
[12] J.R. Jackson, Networks of waiting lines, Oper. Res. 5 (4) (1957) 518–521, https://doi.org/10.1287/opre.5.4.518.
[13] J.R. Jackson, Jobshop-like queueing systems, Manag. Sci. 10 (1) (1963) 131–142, https://doi.org/10.1287/mnsc.10.1.131.
[14] C. Jacobi, K. Furmans, Data sets for the analysis of decomposition error in discrete-time open tandem queues, Repository KITopen, 2022.
[15] J.L. Jain, W.K. Grassmann, Numerical solution for the departure process from the GI/G/1 queue, Comput. Oper. Res. 15 (3) (1988) 293–296, https://doi.org/10.1016/0305-0548(88)90042-1.
[16] S. Kim, R. Muralidharan, C.A. O'Cinneide, Taking account of correlations between streams in queueing network approximations, Queueing Syst. 49 (3) (2005) 261–281, https://doi.org/10.1007/s11134-005-6967-8.
[17] R. Koenker, G. Bassett, Regression quantiles, Econometrica 46 (1) (1978) 33–50.
[18] R. Koenker, V. d'Orey, Algorithm AS 229: computing regression quantiles, J. R. Stat. Soc., Ser. C, Appl. Stat. 36 (3) (1987) 383–393.
[19] R. Koenker, V. d'Orey, Remark AS R92: a remark on algorithm AS 229: computing dual regression quantiles and regression rank scores, J. R. Stat. Soc., Ser. C, Appl. Stat. 43 (2) (1994) 410–414.
[20] R. Koenker, K.F. Hallock, Quantile regressions, J. Econ. Perspect. 4 (15) (2001) 143–156.
[21] R. Koenker, J.A. Machado, Goodness of fit and related inference for quantile regression, J. Am. Stat. Assoc. 94 (1999) 1296–1310.
[22] P. Kuehn, Approximate analysis of general queuing networks by decomposition, IEEE Trans. Commun. 27 (1) (1979) 113–126, https://doi.org/10.1109/TCOM.1979.1094270.
[23] K. Lieckens, N. Vandaele, Multi-level reverse logistics network design under uncertainty, Int. J. Prod. Res. 50 (1) (2012) 23–40, https://doi.org/10.1080/00207543.2011.571442.
[24] M. Reiser, H. Kobayashi, Accuracy of the diffusion approximation for some queuing systems, IBM J. Res. Dev. 18 (2) (1974) 110–124, https://doi.org/10.1147/rd.182.0110.
[25] R. Sagron, D. Grosbard, G. Rabinowitz, I. Tirkel, Approximation of single-class queueing networks with downtime-induced traffic variability, Int. J. Prod. Res. 53 (13) (2015) 3871–3887, https://doi.org/10.1080/00207543.2014.974845.
[26] R. Sagron, G. Rabinowitz, I. Tirkel, Approximating class-departure variability in tandem queues with downtime events: regression-based variability function, Comput. Oper. Res. 88 (2017) 161–174, https://doi.org/10.1016/j.cor.2017.07.003.
[27] M. Schleyer, An analytical method for the calculation of the number of units at the arrival instant in a discrete time G/G/1-queueing system with batch arrivals, OR Spektrum 34 (1) (2010) 293–310, https://doi.org/10.1007/s00291-010-0226-z.
[28] M. Schleyer, K. Furmans, An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals, OR Spektrum 29 (4) (2007) 745–763, https://doi.org/10.1007/s00291-006-0065-0.
[29] M. Schleyer, K. Gue, Throughput time distribution analysis for a one-block warehouse, Transp. Res., Part E, Logist. Transp. Rev. 48 (3) (2012) 652–666, https://doi.org/10.1016/j.tre.2011.10.010.
[30] J.A. Schwarz, M. Epp, Performance evaluation of a transportation-type bulk queue with generally distributed inter-arrival times, Int. J. Prod. Res. 54 (20) (2016) 6251–6264, https://doi.org/10.1080/00207543.2015.1092613.
[31] A.K. Sen, M.S. Srivastava, Regression Analysis: Theory, Methods, and Applications, Springer Texts in Statistics, Springer, New York, 1990.
[32] J.G. Shanthikumar, J.A. Buzacott, Open queueing network models of dynamic job shops, Int. J. Prod. Res. 19 (3) (1981) 255–266, https://doi.org/10.1080/00207548108956652.
[33] J.G. Shanthikumar, S. Ding, M.T. Zhang, Queueing theory for semiconductor manufacturing systems: a survey and open problems, IEEE Trans. Autom. Sci. Eng. 4 (4) (2007) 513–522, https://doi.org/10.1109/TASE.2007.906348.
[34] S. Suresh, W. Whitt, The heavy-traffic bottleneck phenomenon in open queueing networks, Oper. Res. Lett. 9 (6) (1990) 355–362, https://doi.org/10.1016/0167-6377(90)90054-9.
[35] I. Van Nieuwenhuyse, R.B. de Koster, Evaluating order throughput time in 2-block warehouses with time window batching, Int. J. Prod. Econ. 121 (2) (2009) 654–664, https://doi.org/10.1016/j.ijpe.2009.01.013.
[36] W. Whitt, The queueing network analyzer, Bell Syst. Tech. J. 62 (9) (1983) 2779–2815, https://doi.org/10.1002/j.1538-7305.1983.tb03204.x.
[37] W. Whitt, Variability functions for parametric-decomposition approximations of queueing networks, Manag. Sci. 41 (10) (1995) 1704–1715, https://doi.org/10.1287/mnsc.41.10.1704.
[38] R.W. Wolff, Stochastic Modeling and the Theory of Queues, Prentice Hall International Series in Industrial and Systems Engineering, Prentice Hall, 1989.
[39] G. Wu, X. Xu, Y.Y. Gong, R. De Koster, B. Zou, Optimal design and planning for compact automated parking systems, Eur. J. Oper. Res. 273 (3) (2019) 948–967, https://doi.org/10.1016/j.ejor.2018.09.014.
[40] K. Wu, L. McGinnis, Interpolation approximations for queues in series, IIE Trans. 45 (3) (2013) 273–290, https://doi.org/10.1080/0740817X.2012.682699.
[41] M. Yu, R.B. de Koster, The impact of order batching and picking area zoning on order picking system performance, Eur. J. Oper. Res. 198 (2) (2009) 480–490, https://doi.org/10.1016/j.ejor.2008.09.011.