Control Charts for Traffic Intensity Monitoring of Markovian Multiserver Queues[‡]

F. R. B. Cruz[§]
fcruz@est.ufmg.br

R. C. Quinino[§]
roberto@est.ufmg.br

L. L. Ho[¶]
lindalee@usp.br

March 9, 2020

Abstract — **A number of recent research studies have applied queueing theory as an approximate modeling tool to mathematically describe industrial systems, which include manufacturing, distribution, and service, for instance. Among the main observable characteristics in queues, the number of users in the system can be controlled to keep waiting times as minimal as possible. The design of efficient control charts is an attempt to monitor and control such systems. Control charts are proposed to monitor infinite queues with Markovian arrivals, exponential service times, and s identical parallel servers. The proposed charts monitor traffic intensities, which are the ratio between the arrival rate and the service rate, estimated through the number of users in the queueing system at random epochs. The effectiveness and efficiency of the proposed approaches in terms of the average run lengths are established by a comprehensive set of Monte Carlo simulations.**

Keywords — Quality control; attribute control charts; Markovian queues; average run length.

1 Introduction

The use of queueing models has been the subject of a number of research studies [10, 29, 36], mainly due to their ability to approximately represent industrial systems [9, 12, 34]. In general, basic queueing models are applied as approximations for complex computer and telecommunication networks [25, 30], manufacturing and service systems [11, 16, 28, 39], and, more recently, healthcare systems [2, 3, 47], among others. In particular, $M/M/s$ queues are one of the most basic queueing models [18], which, in Kendall notation, stand for Markovian arrivals with rate λ, exponential service times with average $1/\mu$, and s parallel identical servers. Despite their simplicity, these models may find application in real-life systems [11].

Queueing models in general and $M/M/s$ queues in particular are especially useful when predicting performance measures from the systems they model, such as the empty system probability (P_0), expected number of customers in the systems (L), expected number of customers in the queue (L_q), expected time in the system (W), and expected time in the queue (W_q). The first three previous performance measures can be derived from the traffic intensity, defined as the ratio $\rho = \lambda/s\mu$, and ρ can be simply estimated by observing the number of users in the systems at random epochs [11]. Given the importance of the traffic intensity, this paper addresses the challenge of monitoring its values through a control chart.

In other words, the goal is to propose control charts to efficiently monitor changes in traffic intensity ρ. Thus, let us suppose that the process under analysis requires the identification of a reduction or an increase in ρ from ρ_0 to another unknown value $\rho_1 \neq \rho_1$ after a random time, which is equivalent to the following hypothesis testing:

$$\begin{cases} \mathrm{H}_0 : \rho = \rho_0, \\ \mathrm{H}_1 : \rho \neq \rho_0, \end{cases}$$

after a random time. To justify the use of $\rho \neq \rho_0$ as an out-of-control case, it is considered that low and high values of ρ mean, respectively, that a greater or a lower number of servers than necessary are in use, consequently leading to system settings that are too expensive or too slow. That is, ρ_0 is acceptable to customers in terms of short waiting times and is economically viable in terms of the low number of servers s.

The problem of designing a control chart for ρ involves fixing the type I error α, that is, the probability of false alarms that indicate that H_0 should be rejected, determining the lower control limits (LCL) and upper control limits (UCL) such that

$$P(\mathrm{LCL} < \hat{\rho} < \mathrm{UCL}) |\, \rho = \rho_0) = \alpha, \qquad (1)$$

in which $\hat{\rho}$ is some estimate for ρ based on a sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ from the number of customers in

[‡]Quality and Reliability Engineering International. February 2020, Volume 36, Issue 1, p. 354-364. DOI: 10.1002/qre.2578. The final publication is available at https://onlinelibrary.wiley.com.

[§]Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901 - Belo Horizonte - MG, Brazil

[¶]Department of Production Engineering, University of São Paulo, 05508-010 - São Paulo – SP, Brazil.

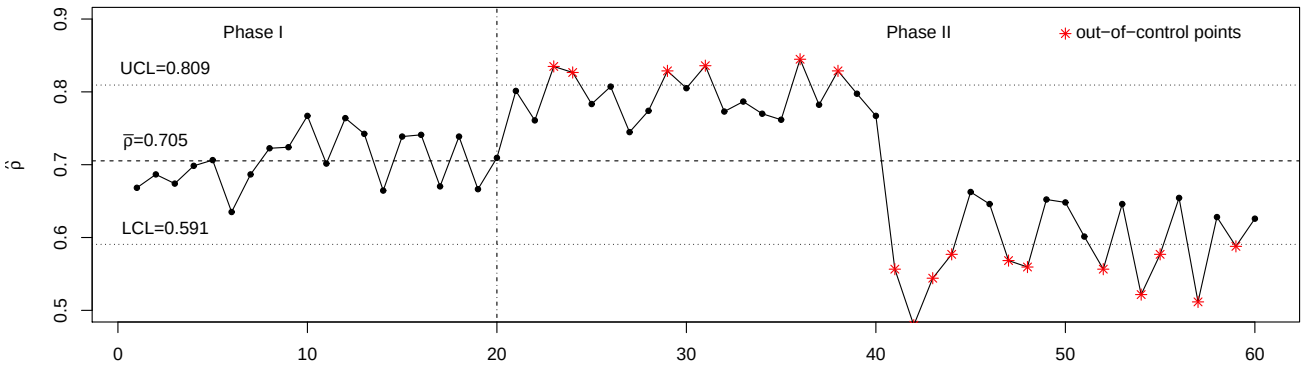


Figure 1: Control chart for ρ (phase I, 1–20; phase II, 21–60).

the $M/M/s$ queueing system at n random epochs. In other words, the control limits LCL and UCL are such that

$$\int_{\text{LCL}}^{\text{UCL}} f(\rho|\text{H}_0)d\rho = \alpha, \tag{2}$$

in which $f(\rho|\text{H}_0)$ is the probability density function of the estimator.

Note that a traditional Shewhart-type 6σ control chart based on the normal distribution of the estimator ρ has a correspondent type I error (false alarm) probability of $\alpha = 0.002699796$ and, consequently, an average length run under H_0, ARL_0, equal to $1/\alpha \approx 370.3983$, as given by the geometric distribution. Additionally, such a control chart would ideally have an average length run under H_1 of $\text{ARL}_1 = 1/(1-\beta)$, with β being the type II error (failure in rejecting H_0) probability, which should be as short as possible. Figure 1 shows a control chart for ρ for simulated data, using $\rho_0 = 0.70$ and type I error $\alpha \approx 0.002699796$ to yield $\text{ARL}_0 \approx 370.3983$. Phase I goes from 1 to 20, that is, data from 1 to 20 are collected at in-control state H_0, and the upper and lower control limits are computed. Then, phase II goes from 21 to 60 with data collected at an out-of-control state, with $\rho = \rho_1 > \rho_0$ for data from 21 to 40 and $\rho = \rho_1 < \rho_0$ for data from 41 to 60.

The problem of developing control charts specifically for queues has existed for quite some time. One of the first times the problem appeared in the literature was in an attempt to control the traffic intensity in general single-server $M/G/1$ and $GI/M/1$ queues [6]. The problem appeared in the literature as an example of a general control chart for attributes when a new control chart for the $M/M/s$ queueing model was developed by Shore [44] for the number of customers in the queueing system (either being served or waiting). Shore later extended this control chart to monitor the queue length in a more general $G/G/s$ queue [45]. Both charts stated their efficiency in terms of the nominal tail areas [44, 45]. Variants were developed some time later to efficiently monitor the queue length in pure Markovian single-server queues, $M/M/1$, and extensions by means of Shewhart-type control charts [23], control charts using the method of weighted variances for the random queue length [13], and a cumulative sum (CUSUM) scheme [7], with the performance of these charts given by the average run lengths (ARL) and the false alarm rate α. Markovian Erlang-serviced single-server queues, $M/E_k/1$, were the object of a study [40] in which control charts were provided to control limits for the random queue length so that customers could have a prior idea about expected waiting times, maximum waiting times, and minimum waiting times based on the central line, the upper control limit, and the lower control limit of the chart so that the performance would be improved. Another extension, Markovian infinite server queues, $M/M/\infty$, was treated in terms of Shewhart-type control charts to control the random queue length [38], with the performance compared using the average run length (ARL) as the performance measure. However, the random queue length is not the only control variable that has been used. Markovian single-server queues were also successfully controlled by means of the traffic intensity [4]. Even sophisticated schemes were proposed to consider data auto-correlation by means of control charts based on the weighted likelihood ratio test (WLRT) [42], for which numerical results and an illustrative example were presented to assess the performance of the proposed WLRT chart as satisfactory.

Although we have chosen to stop here, as many references could easily be added to the above list, to the best of our knowledge, control charts for controlling the traffic intensity in Markovian multi-server queues that consider the problem of estimating the parameters have not been treated in the literature. Thus, we propose control charts to control traffic intensity ρ as a way of controlling ρ itself and indirectly P_0, L, and L_q.

The remainder of this paper is organized as follows. Section 2 details the methods developed. Results from computer simulations and a detailed discussion to support the quality of the proposed approach are presented in Section 3. Finally, Section 4 concludes the paper with final remarks and some topics for future research in this

area.

2 Material and Methods

2.1 *Inference in* M/M/s *Queues*

Fixing the traffic intensity, $0 < \rho < 1$, the $M/M/s$ queueing system has a stationary distribution if it is in equilibrium. From the theory developed for the birth-death process (for instance, see Gross et al. [18]; Ross [43]; or Bhat [5]), the stationary distribution of the number of customers (N) in the system at random epochs is expressed by

$$P_n \equiv P(N = n) = \begin{cases} \dfrac{(s\rho)^n}{n!} P_0, & 0 \leq n < s, \\ \dfrac{s^s \rho^n}{s!} P_0, & n \geq s, \end{cases} \tag{3}$$

in which $P_0 \equiv P(N = 0)$ is given by the usual boundary condition that the probabilities must sum to 1 as follows:

$$P_0 = \left(\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!} \frac{1}{1-\rho} \right)^{-1}. \tag{4}$$

The focus here is on the traffic intensity ρ, which is somewhat in contrast to several existing studies in which inferences are performed on λ and μ. Thus, it is necessary to observe the system at sufficiently spaced random epochs to avoid correlation to generate data ensuring that the data-generating process is consistent with the probability distribution in Eq. (3). Then, assuming that the number of customers found in the queue is x_i and that $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ constitutes our sample of size n, the corresponding likelihood function is

$$L(\mathbf{x}|\rho) = \prod_{i=1}^{n} \left[\frac{(s\rho)^{x_i}}{x_i!} P_0 I_{\{0 \leq x_i < s\}} + \frac{s^s \rho^{x_i}}{s!} P_0 I_{\{x_i \geq s\}} \right], \tag{5}$$

where $I_{\{\bullet\}}$ is the indicator function. Again, to ensure independence of the sample observations, they must be sufficiently spaced in time. The value that maximizes the likelihood function is known as the maximum likelihood estimator (MLE) for ρ; that is,

$$\rho_{\text{MLE}} = \arg\max_{0<\rho<1} L(\mathbf{x}|\rho), \tag{6}$$

in which ρ_{MLE} is the maximum likelihood estimator for ρ. However, because the likelihood function given by Eq. (5) is not algebraically simple, it is necessary to conduct a numerical search to find an estimate for the traffic intensity, $\hat{\rho}_{\text{MLE}}$, given an observed sample $\mathbf{x}$. Among all the different numerical methods that can be used to compute its value, the Golden-section method was chosen because of its efficiency, efficacy, and simplicity [41]. The method is shown in Figure 2. For an in-depth view on estimation in $M/M/s$ queues, see Suyama et al. [46]

algorithm
  `Epsmlemms` $\leftarrow 1.0\text{E}-03$; % accuracy value
  `iter` $\leftarrow 50$; % maximum number of iterations
  $\tau \leftarrow \left(\sqrt{5}-1\right)/2$; % golden proportion coefficient
  $a \leftarrow 0$; $b \leftarrow 1$;
  $c \leftarrow b - \tau(b-a)$;
  $d \leftarrow a + \tau(b-a)$;
  $k \leftarrow 1$;
  while `abs`$(d-c) >$ `Epsmlemms` **and** $k <$ `iter`
    if $L(\mathbf{x}|c) > L(\mathbf{x}|d)$ **then**
      $b \leftarrow d$;
    else
      $a \leftarrow c$;
    endif
    $c \leftarrow b - \tau(b-a)$;
    $d \leftarrow a + \tau(b-a)$;
    $k \leftarrow k + 1$;
  endwhile
  $\hat{\rho}_{\text{MLE}} \leftarrow (a+b)/2$;
end algorithm

Figure 2: Golden section algorithm.

2.2 *Control Charts for* M/M/s *Queues*

In the construction of Shewhart-type control charts to monitor traffic intensity ρ by determining symmetrical control limits based on the approximate normal distribution, the desired value of $\text{ARL}_0 = 1/\alpha \approx 370.3983$ may not be reached depending on the sample size [21, 19, 20]. Thus, different procedures must be proposed. In fact, Figure 3 shows a histogram for 10,000 maximum likelihood estimates for the traffic intensity, $\hat{\rho}_{\text{MLE}}$, for small samples of size $n = 50$ drawn from queueing systems with known $\rho = 0.80$. Such approximation by a normal distribution may not be ideal and may lead to errors. Thus, to determine the upper and lower control limits, UCL and LCL, respectively, we propose methods based on bootstrapping and kernel estimation as follows.

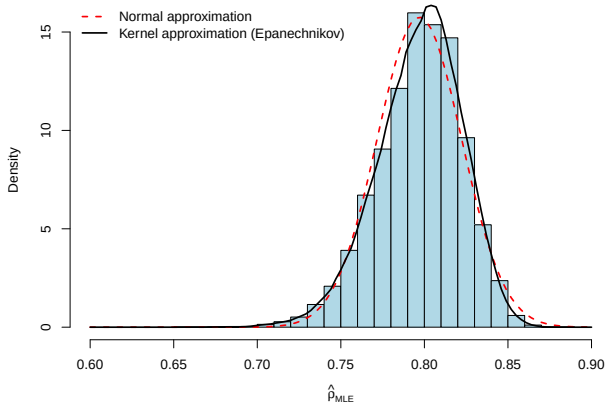


Figure 3: Empirical distribution of $\hat{\rho}_{\text{MLE}}$ for $\rho = 0,80$, and $n = 50$.

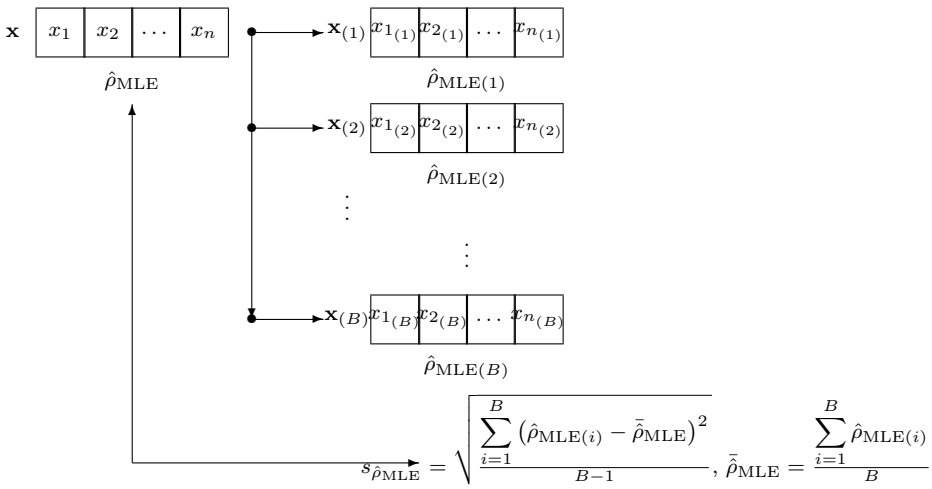


Figure 4: The bootstrap standard deviation method.

2.2.1 *Bootstrap Standard Deviation (BSD) Control Chart*

Proposed by Efron [14], bootstrapping is a well-known computationally intensive technique. The bootstrap scheme is illustrated in Figure 4, where, in its non-parametric version, B re-samplings with replacement $\mathbf{x}_{(i)}$ (typically $B \geq 200$) are drawn from the original sample $\mathbf{x}$ (collected under H_0), and the maximum likelihood estimates of the traffic intensity are obtained for each of them, $\hat{\rho}_{\text{MLE}(i)}$. The bootstrap method can be particularly useful in situations where the distribution of the parameter of interest is unknown [15], as is the case here. In fact, the use of the bootstrap method in control charts goes back to the work of Zhang & Wang [50], where comparisons were made with the traditional Shewhart-type control charts, and the results were encouraging. A close view of bootstrap control charts can be seen in the work of Jones & Woodall [22], where extensive computer simulations were provided to assess the performance of bootstrap control charts in terms of the average run length (ARL). Since then, several bootstrap control charts have been developed with successful results, such as control charts for Weibull percentiles [37], Birnbaum-Saunders percentiles [31], inverse Gaussian percentiles [32], and autocorrelated process data [33] and bootstrap-based maximum multivariate cumulative sum charts [24], among others.

One possible way to determine the upper and lower control limits UCL and LCL, respectively, is by means of the bootstrap standard deviation. The average $\bar{\hat{\rho}}_{\text{MLE}}$ and the standard deviation $s_{\hat{\rho}_{\text{MLE}}}$ of the MLE obtained from the B bootstrap samples are calculated. Considering $\alpha = 0.002699796$, for the traditional 6σ limits, the upper and lower control limits become

$$\begin{cases} \text{UCL}_{\text{BSD}} = \bar{\hat{\rho}}_{\text{MLE}} + 3 \times s_{\hat{\rho}_{\text{MLE}}}, \\ \text{LCL}_{\text{BSD}} = \bar{\hat{\rho}}_{\text{MLE}} - 3 \times s_{\hat{\rho}_{\text{MLE}}}. \end{cases} \tag{7}$$

2.2.2 *Percentile Bootstrap (PB) Control Chart*

Another option considered here is the use of percentiles $(1-\frac{\alpha}{2})\times 100\%$ and $\frac{\alpha}{2}\times 100\%$ of the B bootstrap estimates $\hat{\rho}_{\text{MLE}(\bullet)}$, that is (for a 6σ control chart and respective $\alpha = 0.002699796$),

$$\begin{cases} \text{UCL}_{\text{PB}} = \hat{\rho}_{\text{MLE}(\bullet);(99.86501020)}, \\ \text{LCL}_{\text{PB}} = \hat{\rho}_{\text{MLE}(\bullet);(0.13498980)}. \end{cases} \tag{8}$$

2.2.3 *Kernel-based (KB) Control Chart*

Because the density probability function $f(\rho)$ of the maximum likelihood estimator (represented here for simplicity without the subscript) is unknown, the control limits cannot be calculated from Eq. (2). A possible way to address this difficulty is through a kernel estimator, a tool that can reveal the density behind a sample of m estimates, $\hat{\boldsymbol{\rho}} = \{\hat{\rho}_{(1)}, \hat{\rho}_{(2)}, \ldots, \hat{\rho}_{(m)}\}$, in which $\hat{\rho}_{(i)}$ is a bootstrap maximum likelihood estimate as defined earlier. The classical model is the Parzen-Rosenblatt estimator,

$$\hat{f}_h(\rho) = \frac{1}{mh}\sum_{j=1}^{m} k\left(\frac{\rho - \hat{\rho}_{(j)}}{h}\right), \tag{9}$$

in which $k(x)$ is the kernel function, assumed to be a symmetric probability density function (such as a normal distribution), $\hat{\rho}$ is the maximum likelihood estimator, and h is a smoothing parameter called the bandwidth [48].

Concerning the density function, $k(x)$, although a Gaussian kernel is typical, the specialized kernel proposed by Zhang et al. [51] is used here to overcome the problem of estimating a strictly positive function domain (that is, $0 < \rho < 1$). It is worth mentioning that other alternatives are possible, including recently developed kernels [1, 26, 27], but in the past, Zhang et al.'s method has performed well for inference in queues [12]. The use of this method results in the following estimator

after a data transformation, $g(x) = x + dx^2 + Adx^3$, with $A > 1/3$ and $d = f'(0)/f(0)$:

$$\hat{f}_h(\rho) = \frac{1}{mh}\sum_{j=1}^{m}\left[k_E\left(\frac{\rho - \hat{\rho}_{(j)}}{h}\right) + k_E\left(\frac{\rho + g(\hat{\rho}_{(j)})}{h}\right)\right], \tag{10}$$

in which $k_E(x)$ is the Epanechnikov kernel, $k_E(x) = \frac{3}{4}(1-x^2)I_{\{-1\le x\le 1\}}$, defined in terms of the indicator function $I_{\{\bullet\}}$.

Concerning the optimal window, h_{opt}, its value may be estimated by using the mean integrated squared error (MISE),

$$\text{MISE}_m(h) = \text{E}\left[\int_{-\infty}^{\infty}\left\{\hat{f}_h(\rho) - f(\rho)\right\}^2 dx\right], \tag{11}$$

which is commonly used to evaluate the performance of the kernel estimation of the density function. The optimal window is given by the minimum of Eq. (11), that is, when $\frac{d\text{MISE}(h)}{dh} = 0$, which produces (see the details in Chiu [8], for instance)

$$h_{\text{opt}} = \left[\frac{\int_{-\infty}^{\infty} k^2(x)dx}{\left\{\int_{-\infty}^{\infty} x^2 k(x)dx\right\}^2 \int_{-\infty}^{\infty}\{f''(x)\}^2 dx}\frac{1}{m}\right]^{\frac{1}{5}}. \tag{12}$$

Note that $\int_{-\infty}^{\infty} k^2(x)dx$ is easily computed once a kernel is chosen, $\int_{-\infty}^{\infty} x^2 k(x)dx$ is the variance of the chosen kernel, and $G = \int_{-\infty}^{\infty}\{f''(x)\}^2 dx$ is the only unknown quantity in the right-hand side of Eq. (12). The plug-in method proposed by Chiu [8] obtains a bandwidth estimate by replacing G with the following estimate of G:

$$\hat{G} = \frac{1}{\pi}\int_0^{\Lambda}\lambda^4\left\{|\hat{\phi}(\lambda)|^2 - \frac{1}{m}\right\}d\lambda. \tag{13}$$

In Eq. (13), Λ is the first value of λ such that $|\hat{\phi}(\lambda)|^2 \le c/n$ for some constant $c > 1$ (after some experimentation, it was found that $c = 3$ yields an estimator with the smallest variance), and $\hat{\phi}(\lambda)$ is the sample characteristic function

$$\hat{\phi}(\lambda) = \frac{1}{m}\sum_{j=1}^{m}\exp\left(i\lambda\hat{\rho}_{(j)}\right). \tag{14}$$

From the definition of the sample characteristic function, Eq. (14), it follows that

$$|\hat{\phi}(\lambda)|^2 = \left[\frac{\sum_{j=1}^{m}\cos(\lambda\hat{\rho}_{(j)})}{m}\right]^2 + \left[\frac{\sum_{j=1}^{m}\sin(\lambda\hat{\rho}_{(j)})}{m}\right]^2. \tag{15}$$

After finishing the estimation of the probability density function of the MLE estimator, $\hat{f}_h$, the control limits can be found by means of Eq. (2) and some numerical integration method, considering that the upper control limit (UCL) corresponds to the $\left(1 - \frac{\alpha}{2}\right)\times 100\%$ percentile of $\hat{f}_h$ and the lower control limit (LCL) corresponds to the $\frac{\alpha}{2}\times 100\%$ percentile of $\hat{f}_h$, that is,

$$\begin{cases} \text{UCL}_{\text{KB}} = \left\{\rho^* | \int_{-\infty}^{\rho^*}\hat{f}_h(\rho)d\rho = 1 - \frac{\alpha}{2}\right\}, \\ \text{LCL}_{\text{KB}} = \left\{\rho^* | \int_{-\infty}^{\rho^*}\hat{f}_h(\rho)d\rho = \frac{\alpha}{2}\right\}. \end{cases} \tag{16}$$

3 Numerical Results

An implementation in MATLAB [35] has been developed and used to verify the efficiency of the proposed approaches. The code is available from the authors upon request for educational and research purposes. In the following paragraphs, computational results are presented and discussed.

In production processes, the performance of the control chart is usually measured in terms of the number of samples until a signaling is observed. The most commonly used metric is the average run length (ARL). When a process is in-control, ARL_0 values as large as $1/\alpha \approx 370.3983$ are desirable, where $\alpha = 0.002699796$ is the type I error (false alarm) probability in a 6σ control chart. On the other hand, when the process is out-of-control, small values of ARL_1, given as $1/(1-\beta)$, are preferable, where β is the type II error (failure in rejecting H_0) probability. Table 1 shows the simulation results for the usual 6σ control charts for queueing systems with $s = 4$ servers, three different sample sizes $n = \{50, 100, 200\}$, four different deviations from H_0 $\delta \equiv \rho_1 - \rho_0 = \{-0.10, -0.05, 0.05, 0.10\}$, $B = 100,000$ bootstrap replications, and 100,000 Monte Carlo replicates.

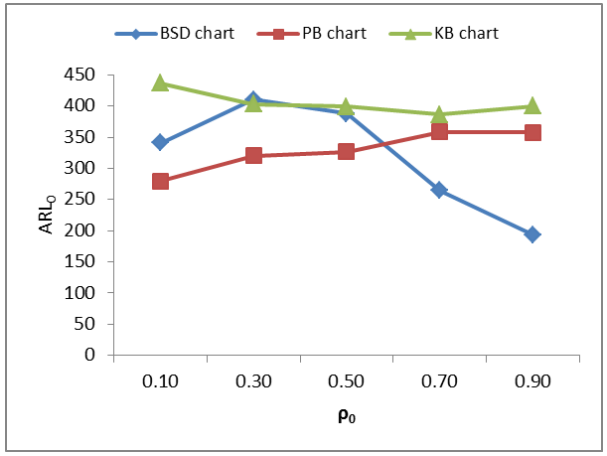
Table 1 shows that the performance of the proposed control charts presents an expected pattern. In fact, under H_0, the average value of ARL_0 is around $1/\alpha \approx 370.3983$. On the other hand, under H_1, the average run length tends rapidly toward 1 as the distance $|\delta| = |\rho_1 - \rho_0|$ increases and the sample size increases.

Table 1 is summarized by Figures 5-(a) to -(f). In general, the values of ARL_0 are the highest on average for the KB control charts. Although sometimes exceeding the nominal value $1/\alpha \approx 370.3983$, the KB control charts provide the longest time between false alarms, as seen in Figures 5-(a) through -(c). Comparing the BSD charts with the PB charts, the latter are preferable under low traffic intensities (i.e., $\rho \le 0.50$) but not under high traffic intensities (that is, $\rho > 0.50$) considering when the PB charts outperform the BSD charts in terms of the ARL_0 mean values, as Figure 5-(a) shows. As the sample size increases, all three control charts present ARL_0 values that go to the nominal value on average, as seen in Figure 5-(b). As expected, the ARL_0 values are on average independent of δ, as presented in Figure 5-(c).

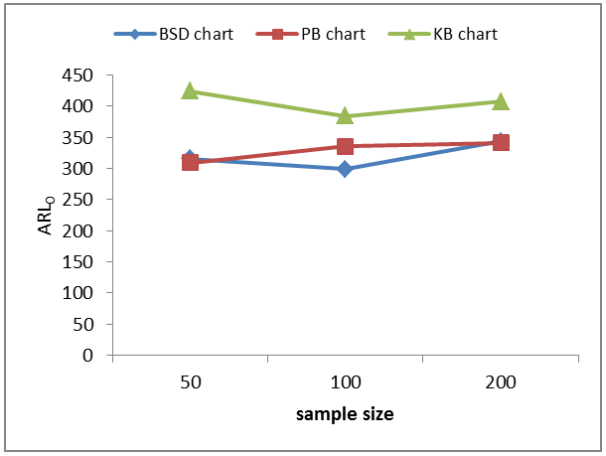
Concerning ARL_1, the PB charts present the lowest mean values and should be preferable if the out-of-control

Table 1: Performance evaluation.

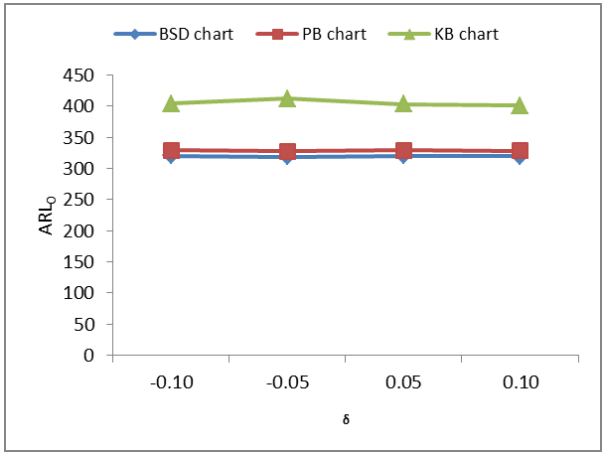
				BSD chart		PB chart		KB chart	
ρ_0	sample size	ρ_1	s	ARL_0	ARL_1	ARL_0	ARL_1	ARL_0	ARL_1
0.10	50	0.01	4	416.67	1.00	245.10	1.00	510.20	1.00
		0.05	4	413.22	7.54	244.50	2.98	510.20	4.51
		0.15	4	411.52	3.90	242.13	3.90	510.20	4.94
		0.20	4	408.16	1.18	243.31	1.18	510.20	1.24
	100	0.01	4	284.90	1.00	296.74	1.00	366.30	1.00
		0.05	4	286.53	1.55	303.95	1.39	374.53	1.39
		0.15	4	284.90	1.76	296.74	1.93	367.65	2.14
		0.20	4	287.36	1.01	303.03	1.01	371.75	1.01
	200	0.01	4	324.68	1.00	295.86	1.00	429.18	1.00
		0.05	4	326.80	1.02	297.62	1.01	432.90	1.01
		0.15	4	325.73	1.12	296.74	1.14	427.35	1.17
		0.20	4	321.54	1.00	293.26	1.00	423.73	1.00
0.30	50	0.20	4	429.18	2.84	262.47	2.11	434.78	2.43
		0.25	4	425.53	31.10	262.47	16.26	436.68	22.26
		0.35	4	432.90	16.30	263.85	16.29	432.90	20.27
		0.40	4	431.03	2.58	265.25	2.58	434.78	2.87
	100	0.20	4	411.52	1.22	350.88	1.18	350.88	1.18
		0.25	4	404.86	8.79	344.83	7.42	384.62	7.42
		0.35	4	401.61	6.94	341.30	6.94	380.23	7.91
		0.40	4	404.86	1.33	344.83	1.33	384.62	1.37
	200	0.20	4	398.41	1.00	354.61	1.00	389.11	1.00
		0.25	4	392.16	2.67	349.65	2.49	434.78	2.67
		0.35	4	395.26	2.77	350.88	2.77	384.62	2.94
		0.40	4	393.70	1.02	349.65	1.02	384.62	1.02
0.50	50	0.40	4	387.60	3.41	331.13	3.89	395.26	4.49
		0.45	4	390.63	23.12	296.74	29.01	400.00	36.87
		0.55	4	390.63	42.75	333.33	19.92	396.83	19.92
		0.60	4	390.63	4.06	333.33	2.82	396.83	2.82
	100	0.40	4	363.64	1.52	318.47	1.58	401.61	1.65
		0.45	4	362.32	9.72	318.47	11.00	401.61	12.47
		0.55	4	371.75	12.24	321.54	8.61	403.23	9.35
		0.60	4	369.00	1.55	318.47	1.40	400.00	1.44
	200	0.40	4	406.50	1.03	334.45	1.04	408.16	1.04
		0.45	4	403.23	3.54	337.84	3.75	408.16	3.98
		0.55	4	406.50	3.89	338.98	3.25	408.16	3.40
		0.60	4	408.16	1.04	340.14	1.03	373.13	1.03
0.70	50	0.60	4	212.31	2.04	347.22	3.00	404.86	3.24
		0.65	4	211.86	10.33	352.11	21.84	411.52	25.18
		0.75	4	213.68	49.85	352.11	8.84	377.36	9.36
		0.80	4	214.13	2.21	348.43	1.35	373.13	1.37
	100	0.60	4	248.76	1.16	362.32	1.30	383.14	1.30
		0.65	4	248.76	4.47	357.14	7.36	398.41	7.92
		0.75	4	246.91	7.38	362.32	3.88	362.32	3.88
		0.80	4	246.91	1.07	362.32	1.03	362.32	1.03
	200	0.60	4	334.45	1.00	362.32	1.01	386.10	1.01
		0.65	4	327.87	2.02	362.32	2.40	390.63	2.46
		0.75	4	334.45	2.10	367.65	1.69	396.83	1.71
		0.80	4	336.70	1.00	361.01	1.00	386.10	1.00
0.90	50	0.80	4	132.63	1.00	369.00	1.01	384.62	1.01
		0.85	4	132.63	1.44	369.00	2.32	384.62	2.27
		0.95	4	133.51	1.16	362.32	1.01	387.60	1.02
		0.99	4	131.93	1.00	355.87	1.00	390.63	1.00
	100	0.80	4	188.68	1.00	346.02	1.00	400.00	1.00
		0.85	4	188.32	1.04	361.01	1.12	387.60	1.12
		0.95	4	189.39	1.00	352.11	1.00	400.00	1.00
		0.99	4	189.39	1.00	349.65	1.00	395.26	1.00
	200	0.80	4	259.07	1.00	355.87	1.00	418.41	1.00
		0.85	4	259.07	1.00	357.14	1.00	420.17	1.00
		0.95	4	259.07	1.00	354.61	1.00	413.22	1.00
		0.99	4	259.07	1.00	358.42	1.00	418.41	1.00



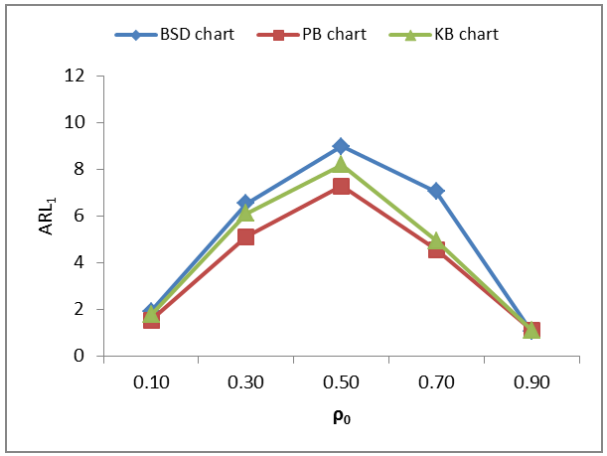
(a) mean ARL_0 *versus* ρ_0



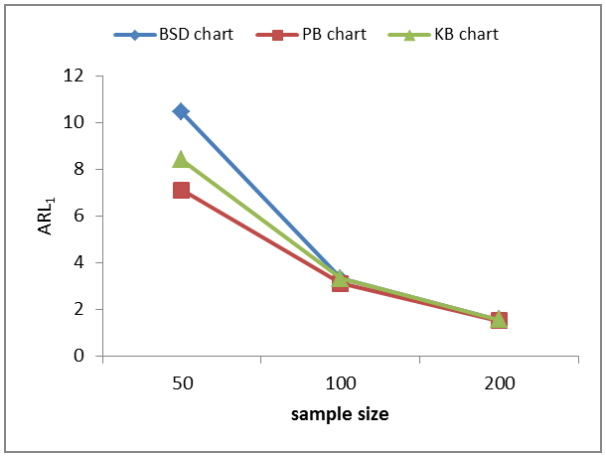
(b) mean ARL_0 *versus* sample size



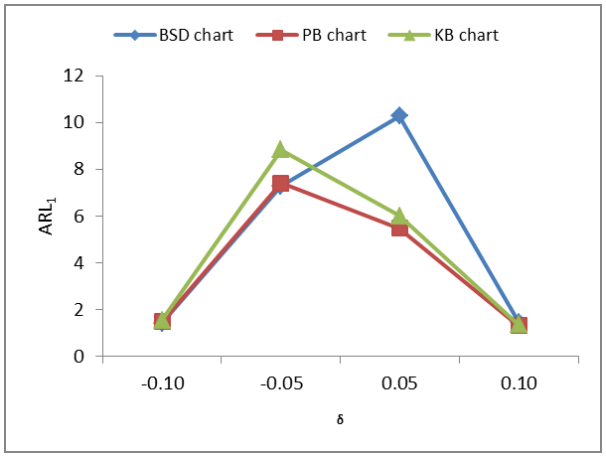
(c) mean ARL_0 *versus* $\delta = \rho_1 - \rho_0$



(d) mean ARL_1 *versus* ρ_0



(e) mean ARL_1 *versus* sample size



(f) mean ARL_1 *versus* $\delta = \rho_1 - \rho_0$

Figure 5: ARL_0 and ARL_1 mean values for the proposed charts.

state prevails in the queueing system. Furthermore, $\rho \approx 0.50$ are the toughest values for the traffic intensity around which shifts out of the in-control state can be identified, as seen in Figure 5-(d), with the highest ARL_1 values on average. Additionally, for extreme traffic intensities ($\rho \to 0.10$ or $\rho \to 0.90$), all three control charts are equally efficient for identifying out-of-control shifts. Additionally, for sample sizes as large as 100 or above, all three control charts present low ARL_1 values on average, as seen in Figure 5-(e), which also shows that small samples of size 50 produce control charts that may be too slow to detect out-of-control shifts. Finally, it is worth mentioning the asymmetrical nature of the density probability function (see Figure 1) that is evident from the results presented in Figure 5-(f). Indeed, the BSD control charts, which assume a normal approximation, tend to have the highest ARL_1 values on average if the traffic intensity increases such that $\rho_1 - \rho_0 \approx 0.05$. On the other hand, the KB charts that attempt to take into account such an asymmetry present the highest ARL_1 values on average when the traffic intensity reduces such that $\rho_1 - \rho_0 \approx -0.05$. More robust than these two charts, the PB control charts exhibit the best performance. In this sense, the KB control charts are preferable if the in-control state prevails in the queueing system, but they are less preferred if the process frequently shifts to an out-of-control state, for which it is better to have a PB chart controlling the system.

4 Conclusions

As previously noted by Green et al. [17], queueing models are important analytical tools to model complex environments. Although such models can never capture all the characteristics of a real operating setting, it has been demonstrated over the years that in a wide range of real situations, queue models can be valuable in providing decision support that is able to significantly improve the performance. In this paper, the problem of controlling the traffic intensity in Markovian multi-server queues ($M/M/s$ queues in Kendal notation), one of the basic queueing models, was approached by means of three original control charts: a chart based on normal approximation and bootstrap standard deviation (BSD control charts), a percentile bootstrap (PB) control chart, and a kernel-based (KB) chart. In the context tested, the control charts produced results that make sense. That is, when the system has an in-control state, high values are obtained for the average run length under H_0 (that is, ARL_0) around the nominal value $1/\alpha \approx 370.3983$ in a 6σ control chart, and when the system is out-of-control, low values close to 1 are obtained for the average run length under H_1 (that is, ARL_1), which is encouraging. Although the computational results seem to suggest the use of KB control charts to monitor $M/M/s$ queueing systems under H_0, these charts may not be ideal for detecting shifts to H_1, for which the PB control charts present the smallest ARL_1 values on average.

There are many directions for further research on this important topic. A first step was taken here in controlling ρ by its upper and lower control limits to ensure the economic viability of a process and quality of service on the user side. Next steps include controlling ρ for more general queueing systems. This would certainly lead to important performance differences because of the differences in the corresponding stationary distributions of the number of customers in the system. Another line of research includes considering variations in the arrival rate λ and a search for control strategies via automatic adjustment of the number of servers c. Support vector machines represent another emerging technique that has begun to bear fruit in the area of control charts [49]. These are only a few topics for future research in this area.

Acknowledgments

Author contributions

All authors, F.R.B.C., R.C.Q., and L.L.H., contributed equally to the design and implementation of the research, to the analysis of the results, and to the final writing of the manuscript.

Financial disclosure

This research is partially supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, grants 305515/2018-7 and 301994/2018-8) and by FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, grant CEX-PPM-00564-17). Part of this research was developed when the first author was visiting the University of New Mexico funded by CNPq.

Conflict of interest

The founders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] Abdous, B. and Kokonendji, C. C. [2009], 'Consistency and asymptotic normality for discrete associated-kernel estimator', *African Diaspora Journal of Mathematics* **8**(2), 63–70.

[2] Almehdawe, E., Jewkes, B. and He, Q.-M. [2013], 'A Markovian queueing model for ambulance offload delays', *European Journal of Operational Research* **226**(3), 602–614.

[3] Almehdawe, E., Jewkes, B. and He, Q.-M. [2016], 'Analysis and optimization of an ambulance offload delay and allocation problem', *Omega* **65**, 148–158.

[4] Bhat, U. N. [1987], 'A sequential technique for the control of traffic intensity in Markovian queues', *Annals of Operations Research* **8**, 151–164.

[5] Bhat, U. N. [2008], *An introduction to queueing theory: Modeling and analysis in applications*, (Statistics for Industry and Technology), Springer, Dordrecht.

[6] Bhat, U. N. and Rao, S. S. [1972], 'A statistical technique for the control of traffic intensity in the queuing systems $M/G/1$ and $GI/M/1$', *Operations Research* **20**, 955–966.

[7] Chen, N. and Zhou, S. [2015], 'CUSUM statistical monitoring of $M/M/1$ queues and extensions', *Technometrics* **57**(2), 245–256.

[8] Chiu, S.-T. [1991], 'Bandwidth selection for kernel density estimation', *The Annals of Statistics* **19**(4), 1883–1905.

[9] Cruz, F. R. B., Almeida, M. A. C., D'Angelo, M. F. S. V. and van Woensel, T. [2018], 'Traffic intensity estimation in finite Markovian queueing systems', *Mathematical Problems in Engineering* **2018**(Article ID 3018758), 1–15.

[10] Cruz, F. R. B., Duarte, A. R. and Souza, G. L. [2018], 'Multi-objective performance improvements of general finite single-server queueing networks', *Journal of Heuristics* **24**(5), 757–781.

[11] Cruz, F. R. B., Quinino, R. C. and Ho, L. L. [2017], 'Bayesian estimation of traffic intensity based on queue length in a multi-server $M/M/s$ queue', *Communications in Statistics - Simulation and Computation* **46**(9), 7319–7331.

[12] Cruz, F. R. B., Santos, M. A. C., Oliveira, F. L. P. and Quinino, R. C. [2018], 'Estimation in a general bulk-arrival Markovian multi-server finite queue', *Operational Research* **(to appear)**, 1–20. (Available on line 06 October 2018). DOI: 10.1007/s12351-018-0433-y

[13] Dhabe, S. D. and Khaparde, M. V. [2011], 'Control charts for random queue length for $(M/M/1)$: $(\infty$/FCFS) queuing model using skewness and power transformation', *Bulletin of Pure and Applied Sciences* **30**(1), 71–83.

[14] Efron, B. [1979], 'Bootstrap methods: Another look at the jackknife', *The Annals of Statistics* **7**, 1–26.

[15] Efron, B. and Tibshirani, R. [1993], *An introduction to the bootstrap*, Chapman & Hall, London, UK.

[16] Govil, M. K. and Fu, M. C. [1999], 'Queueing theory in manufacturing: A survey', *Journal of Manufacturing Systems* **18**(3), 214.

[17] Green, L. V., Soares, J., Giglio, J. F. and Green, R. A. [2006], 'Using queueing theory to increase the effectiveness of emergency department provider staffing', *Academic Emergency Medicine* **13**(1), 61–68.

[18] Gross, D., Shortle, J. F., Thompson, J. M. and Harris, C. M. [2009], *Fundamentals of queueing theory*, 4th edn, Wiley-Interscience, New York, NY.

[19] Ho, L. L. and Costa, A. [2015], 'Attribute charts for monitoring the mean vector of bivariate processes', *Quality and Reliability Engineering International* **31**(4), 683–693.

[20] Ho, L. L., Fernandes, F. H. and Bourguignon, M. [2019], 'Control charts to monitor rates and proportions', *Quality and Reliability Engineering International* **35**(1), 74–83.

[21] Ho, L. L., Quinino, R. C. and Trindade, A. L. G. [2011], 'An np-control chart for inspection errors and repeated classifications', *Quality and Reliability Engineering International* **27**(8), 1087–1093.

[22] Jones, L. A. and Woodall, W. H. [1998], 'The performance of bootstrap control charts', *Journal of Quality Technology* **30**(4), 362–375.

[23] Khaparde, M. V. and Dhabe, S. D. [2010], 'Control charts for random queue length N for $(M/M/1)$: $(\infty$/FCFS) queuing model', *International Journal of Agricultural and Statistics Sciences* **6**(1), 319–334.

[24] Khusna, H., Mashuri, M., Ahsan, M., Suhartono, S. and Prastyo, D. D. [2018], 'Bootstrap-based maximum multivariate CUSUM control chart', *Quality Technology & Quantitative Management* **(to appear)**, 1–23. (Available on line 26 October 2018). DOI: 10.1080/16843703.2018.1535765

[25] Kleinrock, L. [1976], *Queueing systems - Volume I: Theory*, John Wiley & Sons, pp. 21–53.

[26] Kokonendji, C. C. and Kiesse, T. S. [2011], 'Discrete associated kernels method and extensions', *Statistical Methodology* **8**(6), 497–516.

[27] Kokonendji, C. C. and Varron, D. [2016], 'Performance of discrete associated kernel estimators through the total variation distance', *Statistics & Probability Letters* **110**(C), 225–235.

[28] Koole, G. and Mandelbaum, A. [2002], 'Queueing models of call centers: An introduction', *Annals of Operations Research* **113**(1-4), 41–59.

[29] Kose, S. Y. and Kilincci, O. [2015], 'Hybrid approach for buffer allocation in open serial production lines', *Computers & Operations Research* **60**, 67–78.

[30] Lakatos, L., Szeidl, L. and Telek, M. [2013], *Introduction to queueing systems with telecommunication applications*, Springer Science & Business Media, New York, NY.

[31] Lio, Y. L. and Park, C. [2008], 'A bootstrap control chart for Birnbaum-Saunders percentiles', *Quality and Reliability Engineering International* **24**(5), 585–600.

[32] Lio, Y. L. and Park, C. [2010], 'A bootstrap control chart for inverse Gaussian percentiles', *Journal of Statistical Computation and Simulation* **80**(3), 287–299.

[33] Mancenido, M. and Barrios, E. [2012], 'An AR-sieve bootstrap control chart for autocorrelated process data', *Quality and Reliability Engineering International* **28**(4), 387–395.

[34] Martins, H. S., Cruz, F. R. B., Duarte, A. R. and Oliveira, F. L. P. [2019], 'Modeling and optimization of buffers and servers in finite queueing networks', *OPSEARCH* **56**(1), 123–150.

[35] MATLAB [2008], *Version 7.6.0.324 (R2008a)*, The MathWorks Inc., Natick, MA.

[36] Nahas, N. [2017], 'Buffer allocation and preventive maintenance optimization in unreliable production lines', *Journal of Intelligent Manufacturing* **28**(1), 85–93.

[37] Nichols, M. D. and Padgett, W. J. [2006], 'A bootstrap control chart for Weibull percentiles', *Quality and Reliability Engineering International* **22**(2), 141–151.

[38] Pandey, A. [2015], 'Control chart for $(M/M/\infty)$: $(\infty$/FIFO) queuing model', *International Journal of Emerging Trends in Science and Technology* **2**(8), 3064–3070.

[39] Papadopolous, H. T., Heavey, C. and Browne, J. [1993], *Queueing theory in manufacturing systems analysis and design*, Springer Science & Business Media.

[40] Poongodi, T. and Muthulakshmi, S. [2012], 'Random queue length control chart for $(M/Ek/1)$: $(\infty$/FCFS) queueing model', *International Journal of Mathematical Archive* **3**(9), 3340–3344.

[41] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. [2007], *Numerical recipes: The art of scientific computing*, 3rd edn, Cambridge University Press, New York, NY.

[42] Qi, D., Li, Z., Zi, X. and Wang, Z. [2017], 'Weighted likelihood ratio chart for statistical monitoring of queueing systems', *Quality Technology & Quantitative Management* **14**(1), 19–30.

[43] Ross, S. M. [1996], *Stochastic processes*, 2nd edn, John Wiley & Sons, New York, NY.

[44] Shore, H. [2000], 'General control charts for attributes', *IIE Transactions* **32**(12), 1149–1160.

[45] Shore, H. [2006], 'Control charts for the queue length in a $G/G/S$ system', *IIE Transactions* **38**(12), 1117–1130.

[46] Suyama, E., Quinino, R. C. and Cruz, F. R. B. [2018], 'Simple and yet efficient estimators for Markovian multi-server queues', *Mathematical Problems in Engineering* **2018**(Article ID 3280846), 7.

[47] van Brummelen, S. P. J., de Kort, W. L. and van Dijk, N. M. [2015], 'Waiting time computation for blood collection sites', *Operations Research for Health Care* **7**, 70–80.

[48] Wand, M. P. and Jones, M. C. [1995], *Kernel smoothing*, Chapman and Hall/CRC, Boca Raton, FL.

[49] Wang, F.-K., Bizuneh, B. and Cheng, X.-B. [2019], 'One-sided control chart based on support vector machines with differential evolution algorithm', *Quality and Reliability Engineering International* **(to appear)**, 1–12. (Available on line 27 February 2019). DOI: abs/10.1002/qre.2465

[50] Wu, Z. and Wang, Q. [1996], 'Bootstrap control charts', *Quality Engineering* **9**(1), 143–150.

[51] Zhang, S., Karunamuni, R. J. and Jones, M. C. [1999], 'An improved estimator of the density function at the boundary', *Journal of the American Statistical Association* **94**(448), 1231–1240.