Weighted Likelihood Ratio Chart for Statistical Monitoring of Queueing Systems

Dequan Qi[1], Zhonghua Li[2], Xuemin Zi[3], Zhaojun Wang[2*]

[1]LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, China

[2]LPMC and Institute of Statistics, Nankai University, Tianjin 300071, China

[3]School of Science, Tianjin University of Technology and Education, Tianjin 300222, China

Abstract

In recent years, effective monitoring of queueing systems has increasingly attracted attention of researchers in the area of statistical process control (SPC). Most existing works in the literature, however, did not consider the data autocorrelation, nor rigorously evaluate the performance. In this paper, considering the data autocorrelation, a control chart based on the weighted likelihood ratio test (WLRT) is proposed to efficiently monitor the utilization of queueing systems, particularly the M/M/1 queueing system. Our approach can be readily extended to other general queueing systems if the likelihood function can be obtained. Numerical results and illustrative example show that the performance of the proposed WLRT chart is quite satisfactory.

Keywords: Average run length; Queueing systems; Statistical process control; Weighted likelihood ratio statistics.

*Corresponding author. Email: zjwang@nankai.edu.cn

1 Introduction

In the literature, extensive research on queueing systems has been done to estimate the system parameters based on the queueing observations (Clarke, 1957; Ross et al., 2007; Chowdhury and Mukherjee, 2013). In the recent years, statistical monitoring the parameters of queueing systems, such as the utilization, the service rate and the arrival rate, has increasingly attracted attention of researchers. Undoubtedly, such detection methods can be further used to assist root-cause identification and decision making for service-operation improvement. There are, however, several challenges to develop the statistical process control (SPC) methods in the queueing systems (Chen et al., 2011; Chen and Zhou, 2014). 1) The observations collected from queueing systems are only partially available in many situations (Pickands and Stine, 1997). 2) The observations from queueing systems are often auto-correlated (Reynolds, 1968; Hendricks and McClain, 1993). Regrettably, ignoring the date autocorrelation may influence the monitoring performance effectively (Tsung et al., 2006). 3) The distribution of the observations from queueing systems is often highly skewed (Shore, 2006).

The relevant research on this topic can be generally divided into two groups, depending on whether the data autocorrelation is considered. In the first group that ignored the autocorrelation, various control charts have been developed, focusing on the partial sampling scheme (e.g., Bhat and Rao, 1972; Bhat, 1987; Shore, 2006). The partial sampling scheme here implies that we can only observe the queue length Q_n after departure epoch, while the complete sampling scheme implies that we should observe both queue length and system times at arrival and departure epochs. Compared with the complete sampling scheme, the partial sampling scheme is easy and/or inexpensive. Chen et al. (2011) proposed an analytical method based on Markov Chain model to evaluate the efficacy of the WZ chart (Bhat and Rao, 1972) and the nL chart (Shore, 2006). In the second group that considered the autocorrelation, Chen and Zhou (2014) proposed the cumulative sum (CUSUM) scheme to monitor typical queueing systems, in particular the M/M/1 queueing system, for partial sampling scheme and the complete sampling scheme, respectively. Noting that the performance of the CUSUM chart might deteriorate if the real out-of-control (OC) parameters were far from the designated region, Chen and Zhou (2014) suggested using the multiple CUSUM charts with different design parameters or the generalized likelihood ratio (GLR) chart. Nevertheless, the GLR chart cannot be updated recur-

sively and has to be computed by maximizing the likelihood ratio with respect to all possible change locations, which may lead to significant increasing of computational load.

Note that all of the afore-mentioned research ignored the fact that recent data may carry more up-to-date information. Many researchers have shown that giving higher weights to recent data can lead to better monitoring performance, which makes the exponentially weighted moving average (EWMA) charts (e.g., Robert, 1959; Lucas and Saccucci, 1990; Zou and Tsung, 2010; Su et al., 2011; Zi et al., 2013) be widely applied. In general, for EWMA-type control charts, a small value of smoothing parameter leads to better detection of small shifts (e.g., Lucas and Saccucci, 1990; Zhou et al., 2012). Moreover, there is no clear winner between the CUSUM chart and the EWMA chart. For example, Han et al. (2010) compared the performance of the CUSUM and EWMA charts when the observations follow the Poisson distribution and the results showed that the CUSUM charts were superior in dealing with a large shift with a later change in time while the EWMA charts outperformed the CUSUM charts in situations with a small shift and an early change in time.

In this paper, we focus on developing a new control chart based on the weighted likelihood ratio test (WLRT) (Zhou et al., 2012) and comparing it with the CUSUM chart (Chen and Zhou, 2014). Henceforth, the proposed control chart is called WLRT chart for short. Here, we use the WLRT chart rather than the traditional EWMA chart for the following two reasons. On the one hand, Chen and Zhou (2014) have obtained the log-likelihood functions for both the partial and complete sampling schemes in M/M/1 queueing system. Hence, we can extend the log-likelihood functions to the weighted log-likelihood functions and then develop the WLRT chart. On the other hand, the WLRT chart can be readily extended to more general types of queues if we can obtain the likelihood function according to queueing theory. Although we focus on the partial sampling scheme in the M/M/1 queueing system, we can follow the similar procedure for the complete sampling scheme. By our results, we find that, compared with the CUSUM charts, the WLRT chart has more satisfactory IC run length distribution and stands out at early detection.

The rest of this paper is organized as follows. In the next section, the statistical model and the WLRT chart for M/M/1 queueing system are introduced. The following section is devoted to comparing the performance of five methods: WLRT, WZ (Bhat and Rao, 1972), nL (Chen et al., 2011),

CUSUM and GLR (Chen and Zhou, 2014) charts. Finally, an illustrative example and our conclusions are given. The proofs of some properties of the proposed control chart are deferred to the Appendix.

2 The proposed WLRT chart

2.1 WLRT chart

The M/M/1 queueing system is a Poisson-input, exponential-service, single-server queue (Gross and Harris, 1998). We use λ, μ and ρ to denote, respectively, the arrival rate, the service rate and the utilization, where $\rho = \frac{\lambda}{\mu}$. We suppose ρ changes from ρ_0 to another unknown value ρ_1 immediately after an unknown departure epoch τ, which suffices to test the following hypotheses

$$\begin{cases} H_0 : \rho = \rho_0, \\ H_1 : \rho \neq \rho_0, \end{cases}$$

after each departure epoch.

Since the queue lengths Q_{n-1} and Q_n are dependent due to the queueing dynamic, we observe the number of arrivals during the n^{th} service period

$$A_n = Q_n - Q_{n-1} + 1 - Z_{n-1},$$

where Z_{n-1} is an indicator variable which equals 1 if $Q_{n-1} = 0$ and equals 0 otherwise. According to queueing theory, $A_n's$ are independent and identically distributed (iid) variables and

$$Pr\{A_n = k\} = \frac{1}{1+\rho} \cdot (\frac{\rho}{1+\rho})^k, \ k = 0, 1 \ldots.$$

After any departure epoch N, the weighted-log-likelihood function can be derived as

$$l_N(\rho) = \ln \rho \cdot \sum_{n=0}^{N} w_n A_n - \ln(1+\rho) \cdot \sum_{n=0}^{N} w_n (A_n + 1), \tag{2.1}$$

where the weights $w_0 = (1-\theta)^N$, $w_n = \theta(1-\theta)^{N-n}, n = 1, \cdots N$, such that $\sum_{n=0}^{N} w_n = 1$, which are similar to those in Qiu et al. (2010) and Zhou et al. (2012), and $\theta \in (0, 1)$ is a smoothing parameter. Including w_0 and A_0

in equation (2.1) has its own merit, because A_0 can be viewed as a pseudo "sample", and is chosen as ρ_0 here, as $E(A_n) = \rho_0$ under the null hypothesis.

Given the value of θ, we can obtain the maximum weighted likelihood estimate (MWLE) of ρ

$$\hat{\rho}_N = \arg\max_{\rho} l_N(\rho) = \sum_{n=0}^{N} w_n A_n. \tag{2.2}$$

Furthermore, we can express the WLRT statistic as

$$W_N = 2[l_N(\hat{\rho}_N) - l_N(\rho_0)] = 2[\hat{\rho}_N \cdot \ln \frac{\hat{\rho}_N(1+\rho_0)}{\rho_0(1+\hat{\rho}_N)} - \ln \frac{1+\hat{\rho}_N}{1+\rho_0}]. \tag{2.3}$$

When the WLRT statistic in (2.3) is larger than a prespecified upper control limit (UCL), we can declare the system utilization ρ has deviated from the nominal value, which means the system is OC.

In practice, decreases of the service rate and/or increases of the arrival rate are of most interests. Thus, a one-sided WLRT$^+$ chart is desirable. For this purpose, we can develop a one-sided chart for the hypotheses

$$\begin{cases} H_0 : \rho = \rho_0, \\ H_1 : \rho > \rho_0. \end{cases}$$

Following Zhou et al. (2012), by substituting $\tilde{\rho}_N = \hat{\rho}_N I(\hat{\rho}_N > \rho_0) + \rho_0 I(\hat{\rho}_N \leq \rho_0)$ (Shu et al., 2012) into (2.3), the monitoring statistic can be modified by

$$W_N^+ = W_N I(\hat{\rho}_N > \rho_0),$$

when W_N^+ is larger than a UCL, the corresponding control chart generates OC signal.

2.2 Properties of WLRT chart

By some simple algebra (see the Appendix), we get the following properties immediately.

P1. $\hat{\rho}_N$ can be updated recursively

$$\hat{\rho}_N = \hat{\rho}_{N-1} \cdot (1-\theta) + \theta \cdot A_N, \tag{2.4}$$

where the initial value is $\hat{\rho}_0 = \rho_0$ based on A_0 and w_0 defined above.

P2.

$$E(\hat{\rho}_N) = \rho, Var(\hat{\rho}_N) = (\rho + \rho^2) \sum_{n=1}^{N} w_n^2. \tag{2.5}$$

P3. Under null hypothesis, we have

$$\frac{\hat{\rho}_N - \rho_0}{\sqrt{(\rho_0 + \rho_0^2) \sum_{n=1}^{N} w_n^2}} \to^d N(0,1), \tag{2.6}$$

as $\theta N \to \infty$ and $\theta \to 0$.

P4. WLRT statistic $W_N < UCL$ (the system is in-control, IC) is essentially equivalent to

$$a < \hat{\rho}_N < b, \tag{2.7}$$

where a, b $(a < \rho_0 < b)$ are the real roots of the equation $W_N = UCL$.

The property P1 ensures that the computational load will decrease significantly for our WLRT chart due to the recursive representation. And the property P4 makes our proposed WLRT chart look like the traditional EWMA chart, because $\hat{\rho}_N$ admits the classical EWMA updating formulas.

3 Performance comparisons

In this section, we demonstrate the effectiveness of our approach through Monte Carlo simulations (Li et al, 2014). The IC run length distribution, the "true" detection capability, the average run length (ARL), the average number of samples (ANOS) and the relative mean index (RMI) are five criteria used for the performance comparison. Here, the IC run length distribution can be considered satisfactory if it is close to the geometric distribution (Hawkins and Olwell, 1998; Zhou et al., 2012). The "true" detection capability of a chart is reflected by the quantity γ_N, where

$$\gamma_N = Pr_{OC}(RL \leq N) - Pr_{IC}(RL \leq N).$$

A control chart with a larger value of γ_N is considered better (Zhou et al., 2012). The ARL is the average number of points that must be plotted before a point indicates an OC condition (Montgomery, 2013). In comparison of various candidate control charts, ARL or ANOS is very important and also

popular used criterion. When the process is IC, a chart with a larger IC ARL (termed ARL_0) or ANOS (termed ANOS_0) indicates a lower false alarm rate than other charts. When the process is OC, a chart with a smaller OC ARL (termed ARL_1) or ANOS (termed ANOS_1) indicates a better detection ability of process shifts than other charts. The RMI (Han and Tsung, 2006) values can be considered as the average of all relative efficiency values, and a control chart with a smaller RMI value is considered better in its overall performance (Zhou et al., 2012).

As mentioned by some researchers (Borror et al., 1998; Chen et al., 2011; Qiu and Li, 2011; Zhou et al., 2012), a critical issue is whether it is possible and straightforward to find design parameters that ensure the specified IC performance when the data are discrete. By simulation, we find the WLRT^+ chart's ARL_0 can always be attained quite closely if $\theta \leq 0.2$. In order to simplify the UCL calculation for practitioners, the simulated UCLs of the WLRT^+ chart for different smoothing parameters, obtained from 100,000 replications are reported in Table 1. A Fortran program is also available from the authors upon request.

In the scenario with $ARL_0 = 370$ and $\rho_0 = 0.5$, the following simulation is applied to determine the control limit of WLRT^+ chart. The control limits of WLRT^+ chart in other scenarios follow the similar way. To generate the simulated data, we suppose $Q_0 = 0$, which indicates the monitored system starts from an empty queue, and which is common in practice (Chen et al., 2011). We assume further that the arrival rate $\lambda = 0.5$ and the service rate $\mu = 1.0$. Given the UCL, we generate random observation Q_n's and calculate the monitor statistic W_N^+'s until $W_N^+ > UCL$. Then, we obtain the run length (RL). We repeat this procedure 100,000 times and record the values of RL in each repetition to use the bisection searching algorithms to find the control limit such that ARL_0 is about 370.

Hereafter, We use the notation h to denote the control limit coefficients, and obtain all results in this section based on 100,000 replications. For a relatively fair comparison, we adjust the control limits of different charts to make their ARL_0 or ANOS_0 as close as possible. We first compare the one-sided WLRT^+ chart with nL, WZ and CUSUM charts under the assumption that the process change occurs at the same time as the monitoring starts. Two scenarios with the IC utilization $\rho_0 = 0.5$ and $\rho_0 = 0.7$ are considered.

In the scenario with $\rho_0 = 0.5$, we only compare the WLRT^+ chart with CUSUM chart because Chen and Zhou (2014) have revealed that the CUSUM chart outperforms the nL and WZ charts in this scenario. The comparisons

Table 1: Simulated UCL values of WLRT$^+$ chart

	ARL_0							
ρ_0	200	300	370	500	800	1000	2000	5000
				$\theta = 0.025$				
0.3	0.02511	0.03551	0.04123	0.05001	0.06406	0.07089	0.09228	0.12057
0.5	0.02498	0.03539	0.04126	0.04979	0.06394	0.07089	0.09193	0.11996
0.7	0.02494	0.03523	0.04104	0.04961	0.06344	0.07011	0.09163	0.11968
0.9	0.02504	0.03520	0.04091	0.04950	0.06362	0.07047	0.09176	0.11985
				$\theta = 0.05$				
0.3	0.08285	0.10679	0.11950	0.13824	0.16775	0.18185	0.22569	0.28293
0.5	0.08247	0.10634	0.11898	0.13743	0.16645	0.18008	0.22299	0.28000
0.7	0.08177	0.10570	0.11852	0.13716	0.16648	0.17998	0.22368	0.28088
0.9	0.08257	0.10578	0.11821	0.13678	0.16605	0.18003	0.22350	0.28060
				$\theta = 0.1$				
0.3	0.23917	0.28966	0.31567	0.35388	0.41470	0.44349	0.53398	0.65111
0.5	0.23400	0.28481	0.31181	0.35114	0.41263	0.44172	0.53134	0.65077
0.7	0.23686	0.28688	0.31347	0.35182	0.41176	0.44058	0.53105	0.64902
0.9	0.23613	0.28597	0.31217	0.35066	0.41034	0.43936	0.52908	0.64858
				$\theta = 0.2$				
0.3	0.62125	0.72720	0.78392	0.86489	0.99707	1.05945	1.25211	1.50780
0.5	0.62429	0.73068	0.78665	0.86785	0.99454	1.05488	1.24454	1.49919
0.7	0.62190	0.72730	0.78274	0.86331	0.99154	1.05259	1.24511	1.49847
0.9	0.62130	0.72865	0.78339	0.86342	0.99059	1.05164	1.24067	1.49629

of the IC run length distribution and the "true" detection capability when $N \leq 100$ are shown in Figure 1 and Figure 2, respectively. It is obvious from Figure 1 that the IC run length distribution of our proposed WLRT$^+$ chart is satisfactory compared with the CUSUM chart because the distribution of WLRT$^+$ chart is more close to the geometric distribution. Figure 2 reveals that WLRT0.025$^+$ stands out at early detection. In addition, the comparisons of ARL are reported in Table 2. The exact values of ARL_0 are listed in the first row in Table 2, and the corresponding ARL_1 for different shifts in the utilization are summarized in the rest of Table 2. From Table 2, we can observe that the performance of the CUSUM chart might deteriorate if the real OC parameter is far from the design parameter ρ_d. For instance, when $\rho = 0.63$, the ARL_1 of the CUSUM0.6 chart is 97.7, while it is 107 for the CUSUM0.99 chart. We can also find that the performance of WLRT$^+$ charts depends on the smoothing parameter, i.e., charts with smaller parameter θ perform better for detecting small shifts, while those with larger parameter θ perform better for detecting larger shifts. Additionally, the WLRT$^+$ charts perform slightly better at detecting large shifts compared with the CUSUM charts. To evaluate overall performance, we also compute the RMI values in Table 2. Considering the overall performance, WLRT0.025$^+$ outperforms other competitors.

In the scenario with $\rho_0 = 0.7$, we compare the WLRT$^+$ chart with nL, WZ and CUSUM charts. For the nL chart, nonoverlapping sample sums with the sample size n are monitored. The WZ chart generates OC signal when the $A_n's$ are consecutively larger than h for d_u observations. Note that the control limits of the WZ charts are similar to those in Chen et al. (2011), but different from those reported in Bhat and Rao (1972). For the illustration purpose, we only present the comparisons of the ANOS in Table 3. The first row in Table 3 are the exact values of ANOS_0. From Table 3, it can be seen that the performance of CUSUM chart is better than the nL and WZ charts. This result is consistent with the findings by Chen and Zhou (2014). Moreover, the performance of WLRT0.025$^+$ is satisfactory compared with other alternative methods.

Finally, we compare the two-sided WLRT chart with GLR chart. Here, we modify the GLR chart proposed by Chen and Zhou (2014) with $\tilde{\rho}_j^k = \frac{1}{j-k+2}$ when $\tilde{\rho}_j^k = 0$, where $\tilde{\rho}_j^k$ is the maximum likelihood estimator of the utilization given the observations from the jth departure to the kth departure. Considering the performance when detect small downward shift, we choose

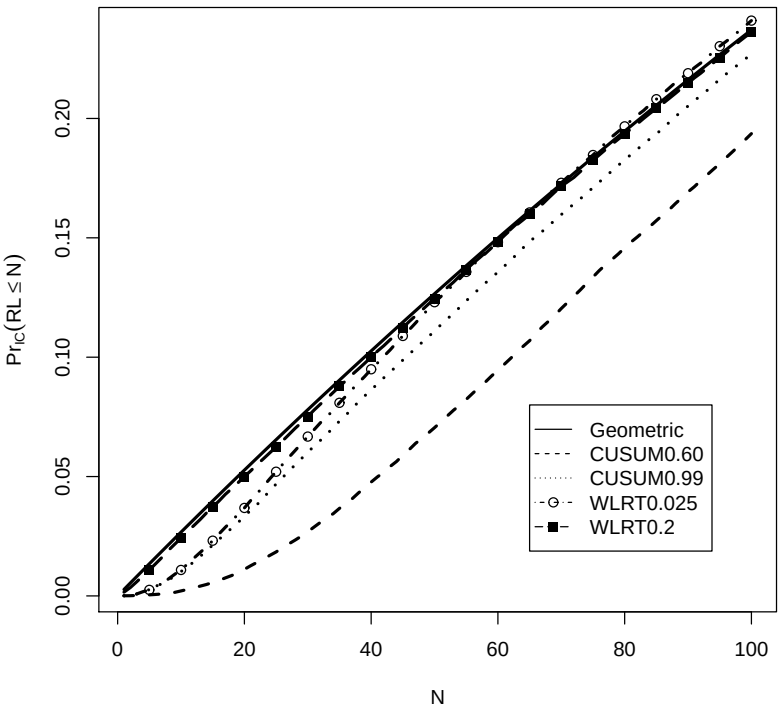

Figure 1: The in-control CDF curves along with Geometric distribution (with expectation 370).

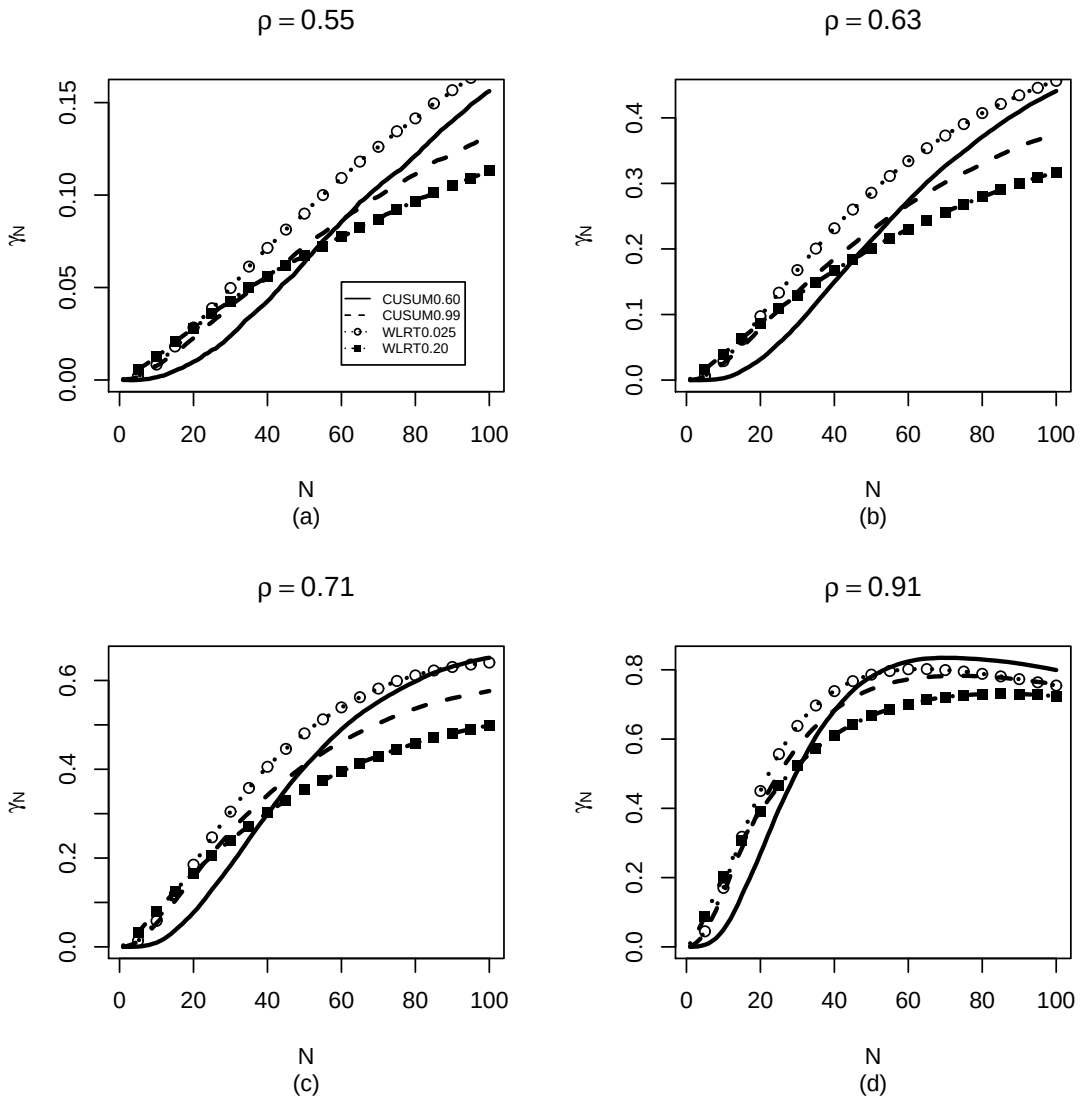

Figure 2: The "true" detection capability between CUSUM and WLRT$^+$. The legend in the first plot is applicable for all others.

Table 2: Comparisons of ARL_1 when $\rho_0 = 0.5$

	CUSUM				$WLRT^+$			
ρ	$\rho_d = 0.6$	0.7	0.85	0.99	$\theta = 0.025$	0.05	0.1	0.2
	$h = 1.38597$	2.09029	2.66849	2.99109	0.04126	0.11898	0.31181	0.78665
0.50	370(339)	369(351)	368(356)	369(362)	370(369)	370(367)	370(368)	370(370)
0.51	320(290)	323(303)	324(311)	327(317)	316(313)	322(318)	329(326)	335(333)
0.55	195(167)	200(181)	207(195)	214(204)	183(175)	197(190)	212(208)	231(230)
0.58	145(118)	148(129)	154(142)	161(152)	132(122)	144(137)	161(157)	180(178)
0.63	97.7(72.9)	98.3(81.2)	103(90.8)	107(98.2)	85.4(75.5)	94.1(86.7)	107(103)	124(122)
0.71	63.1(42.1)	61.5(46.1)	62.6(51.7)	64.9(56.4)	52.4(42.6)	56.8(49.3)	64.5(59.8)	76.2(73.8)
0.91	32.8(18.8)	30.3(19.5)	29.3(21.0)	29.4(22.5)	25.5(18.0)	26.1(20.1)	28.2(24.0)	33.0(30.5)
1.70	11.8(5.91)	10.4(5.67)	9.44(5.67)	9.01(5.70)	8.62(5.23)	8.27(5.38)	8.03(5.72)	8.21(6.54)
3.00	6.13(3.17)	5.48(2.97)	4.87(2.85)	4.60(2.76)	4.53(2.65)	4.28(2.64)	4.06(2.64)	3.91(2.77)
10.0	2.34(1.28)	2.18(1.21)	1.99(1.12)	1.94(1.07)	1.94(1.06)	1.85(1.03)	1.76(0.99)	1.67(0.96)
30.0	1.44(0.68)	1.37(0.64)	1.31(0.58)	1.30(0.56)	1.30(0.56)	1.27(0.54)	1.24(0.51)	1.21(0.48)
RMI	0.244	0.188	0.157	0.160	0.047	0.068	0.113	0.191

[1] NOTE: Standard deviations are in parentheses.

Table 3: Comparisons of $ANOS_1$ when $\rho_0 = 0.7$

	nL		WZ		CUSUM		$WLRT^+$	
ρ	$n = 5$	10	$d_u = 5$	15	$\rho_d = 0.75$	0.98	$\theta = 0.025$	0.1
	$h = 40$	70	7	4	0.68769	2.20254	0.04104	0.31347
0.70	368	372	365	367	370	370	370	370
0.72	309	311	306	307	294	300	290	309
0.76	222	223	220	220	199	206	187	218
0.80	165	167	165	165	146	149	133	162
0.86	114	117	113	116	103	102	87.9	110
0.91	87.9	90.5	87.2	90.5	81.7	78.3	66.9	83.6
0.99	61.8	64.9	61.6	65.7	60.9	55.7	47.4	57.3
1.31	26.8	30.2	26.6	32.6	30.3	25.1	21.1	22.6
1.64	17.5	21.1	17.4	24.4	20.3	16.2	13.6	13.4
1.99	13.4	17.0	13.3	20.9	15.1	11.8	9.95	9.38
2.71	9.87	13.1	9.89	18.3	10.1	7.84	6.62	6.00

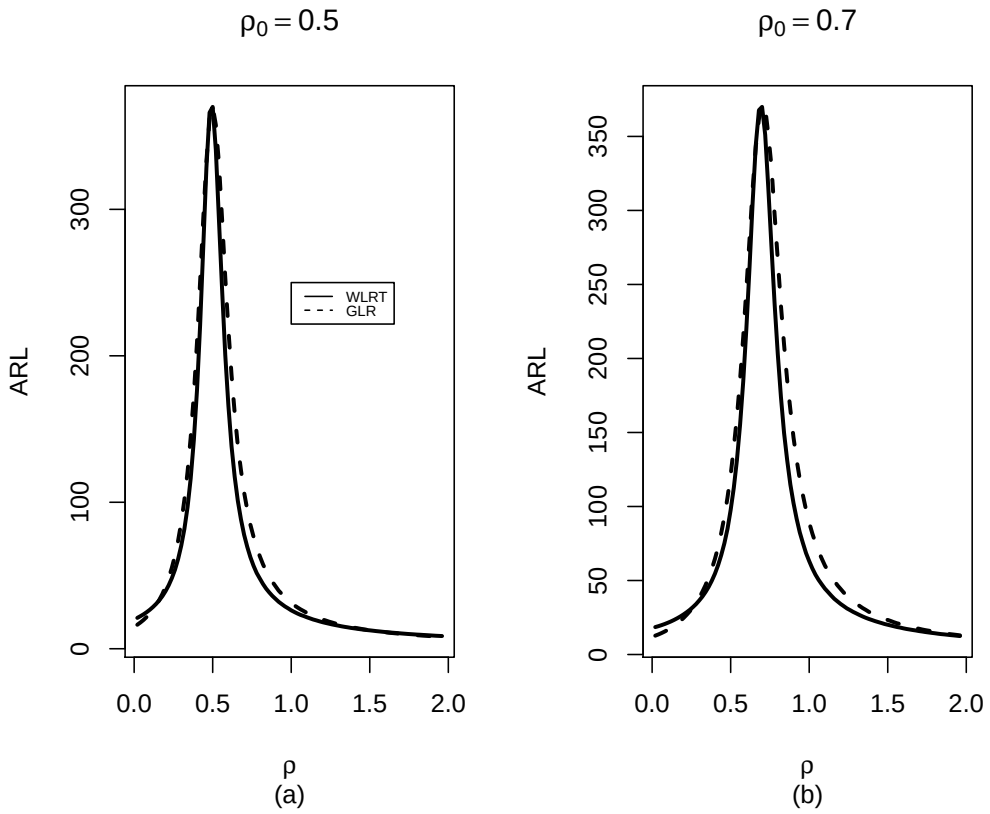

Figure 3: The ARL_1 comparison between WLRT and GLR: (a) $h_{WLRT} = 0.06179$, $h_{GLR} = 4.86283$; (b) $h_{WLRT} = 0.06158$, $h_{GLR} = 4.95644$. The legend in the plot (a) is applicable for the plot (b).

the smoothing parameter θ of the WLRT chart as 0.025. We adjust their control limits such that the ARL_0 is around 370 by convention. We suppose that only the service rate μ changes in different magnitudes which causes the system utilization ρ in both scenarios to shift from 0.02 to 1.96. The corresponding ARL_1 are compared in Figure 3. From Figure 3, we find that there is no evident difference between these two charts in their ability to detect small downward shifts in the utilization. Furthermore, the WLRT0.025 chart performs worse at detecting large downward shifts, but performs better at detecting medium upward shifts.

4 Illustrative Example

In this section, we change the M/G/1 make-to-order production plant model proposed by Chen and Zhou (2014) into M/M/1 as an illustrative example. The orders arrive according to a Poisson process with rate $\lambda = 2.0$ per day. Each order needs to be transacted in an integrated machine center, and only one order can be processed at one time. The service times are independent and exponentially distributed random variables with rate $\mu = 3.0$, and the utilization rate when the operation is normal is therefore $\rho = 0.67$. As mentioned earlier, in the partial sampling scheme, we only observe queue length after departure epoch, and do not know the system times. We consider the service rate changes from 1/8 per hour to 1/10.8 per hour after departure epoch τ, which means ρ changes from 0.67 to 0.9 correspondingly.

For the illustration purpose, we only compare the WLRT$^+$ chart with the CUSUM chart. The smoothing parameter θ of the WLRT$^+$ chart is chosen as 0.025. The design parameter ρ_d of the CUSUM chart is chosen as 0.9 because the corresponding CUSUM chart has the smallest ARL_1. We compare the conditional expected delay (CED) (Kenett and Zacks,1998; Lee and Jun, 2012) due to the detection ability being dependent on the time point of the change (Sonesson and Bock, 2003). We discard any series in which a signal occurs before the $(\tau+1)^{th}$ observation. The CED comparison results of the WLRT$^+$ and CUSUM charts are given in Table 4. It is clear that the performance of the WLRT$^+$ chart is satisfactory, especially when the change time τ is early.

Table 4: The CEDs of the control charts

τ	5	10	50	100	200	300	500	1000	ARL_0	UCL
CUSUM	65.9	64.2	60.8	60.7	60.6	60.6	60.3	60.6	369	2.04603
WLRT$^+$	58.0	57.9	59.4	60.6	60.5	60.8	60.4	60.8	370	0.04090

5 Concluding Remarks

In this paper, we propose a control chart for monitoring the M/M/1 queueing system. The proposed chart, termed the weighted likelihood ratio test

(WLRT) chart, is essentially based on calculating the weighted log-likelihood ratio test statistics. The proposed WLRT chart is compared with some existing charts, such as CUSUM, nL and WZ charts, based on the average run length (ARL) and average number of samples (ANOS). Moreover, the WLRT charts have more satisfactory in-control (IC) run length distribution and stand out at early detection.

We focus on the partial sampling scheme not only because the observations from queueing systems are only partially available in many situations, but also we can follow the similar procedure in the complete sampling scheme. The proposed WLRT chart can be readily extended to more general types of queues, e.g, M/G/1, only if we can obtain the likelihood function according to queueing theory, which will be investigated by the authors in the near future. Future research may also include a self-starting version of the new chart (Li et al., 2010), which is not immediate because the transformation in Li et al (2010) is not easy to derive for the queueing systems.

Appendix

In this Appendix, we sketch the proofs of the properties in Section 2.

1. Apparently, $\hat{\rho}_N$ can be updated recursively.

2. According to queueing theory, we have the following

$$\begin{aligned} E(A_n) &= E(E(A_n|T_n)) = \rho, \\ Var(A_n) &= Var(E(A_n|T_n)) + E(Var(A_n|T_n)) = \rho^2 + \rho, \end{aligned}$$

   where T_n is the service time corresponding to the n^{th} departure. Thus, we obtain the property (2.5) immediately.

3. It is not difficult to see that $\frac{\hat{\rho}_N - \rho_0}{\sqrt{(\rho_0+\rho_0^2)\sum_{n=1}^N w_n^2}}$ can be expressed as a linear combination of iid variables, say

$$\frac{\hat{\rho}_N - \rho_0}{\sqrt{(\rho_0 + \rho_0^2)\sum_{n=1}^N w_n^2}} = \frac{\sum_{n=1}^N w_n(A_n - \rho_0)}{\sigma\sqrt{\sum_{n=1}^N w_n^2}},$$

   where $\sigma^2 = Var(A_n) = \rho_0 + \rho_0^2$. This, together with

$$\max_{1\le n\le N} \frac{w_n^2}{\sum_{n=1}^N w_n^2} = \frac{\theta^2}{\sum_{n=1}^N w_n^2} \to 0,$$

give the property (2.6) by the Hajek-Sidak's Theorem.

4. By some simple algebra we get

$$\frac{\partial WLRT_N}{\partial \hat{\rho}_N} = 2 \cdot \ln \frac{\hat{\rho}_N(1+\rho_0)}{\rho_0(1+\hat{\rho}_N)} = 2 \cdot \ln \frac{\hat{\rho}_N + \rho_0\hat{\rho}_N}{\rho_0 + \rho_0\hat{\rho}_N}.$$

It is easy to see that WLRT_N is monotonically increasing (decreasing) on the right (left) side of ρ_0. This completes the proof (2.7).

Acknowledgements

The authors would like to thank Dr. Nan Chen of National University of Singapore for discussions. The authors would also like to thank the Editor and two Referees for their insightful comments. This paper is supported by the National Natural Science Foundation of China Grants 11431006, 11371202, 11131002, 11201246, 11101306, 11271205, 11101198, 11401573, the RFDP of China Grant 20110031110002.

References

[1] Bhat, U. N. (1987). A Sequential Technique for the Control of Traffic Intensity in Markovian Queues. *Annals of Operations Research,* 8, 151-164.

[2] Bhat, U. N. and Rao, S. S. (1972). A Statistical Technique for the Control of Traffic Intensity in the Queueing Systems M/G/1 and GI/M/1. *Operations Research,* 20, 955-966.

[3] Borror, C. M., Champ, C. W. and Rigdon, S. E. (1998). Poisson EWMA Control Charts. *Journal of Quality Technology,* 30, 352-361.

[4] Chen, N., Yuan, Y. and Zhou, S. (2011). Performance Analysis of Queue Length Monitoring of M/G/1 Systems. *Naval Research Logistics,* 58, 782-794.

[5] Chen, N. and Zhou, S. (2014). CUSUM Statistical Monitoring of M/M/1 Queues and Extensions. *Technometrics,* DOI:10.1080/00401706.2014.923787, available at http://amstat.tandfonline.com/doi/full/10.1080/00401706.2014.923787#.U9W5o9qS1dg.

[6] Chowdhury S. and Mukherjee S. P. (2013). Estimation of Traffic Intensity Based on Queue Length in a Single M/M/1 Queue. *Communications in Statistics-Theory and Methods,* 42, 2376-2390.

[7] Clarke, A. B. (1957). Maximum Likelihood Estimates in a Simple Queue. *The Annals of Mathematical Statistics,* 28, 1036-1040.

[8] Gross, D. and Harris, C. M. (1998). *Fundamentals of Queueing Theory.* Inc: John Wiley & Sons.

[9] Han, D. and Tsung, F. (2006). A Reference-Free Cuscore Chart for Dynamic Mean Change Detection and a Unified Framework for Charting Performance Comparison. *Journal of the American Statistical Association,* 101, 368-386.

[10] Han, S. W., Tsui, K. L., Ariyajunya, B. and Kim, S. B. (2010). A Comparison of CUSUM, EWMA, and Temporal Scan Statistics for Detection of Increases in Poisson Rates. *Quality and Reliability Engineering International,* 26, 279-289.

[11] Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement.* New York: Springer-Verlag.

[12] Hendricks, K. B. and McClain, J. O. (1993). The Output Process of Serial Production Lines of General Machines with Finite Buffers. *Management Science,* 39, 1194-1201.

[13] Kenett, R. S. and Zacks, S. (1998). *Modern Industrial Statistics: Design and Control of Quality and Reliability.* Pacific Grove, CA: Duxbury.

[14] Lee, S. H. and Jun, C. H. (2012). A Process Monitoring Scheme Controlling False Discovery Rate. *Communications in Statistics-Simulation and Computation,* 41, 1912-1920.

[15] Li, Z., Zhang, J. and Wang, Z. (2010). Self-Starting Control Chart for Simultaneously Monitoring Process Mean and Variance. *International Journal of Production Research,* 48 (15), 4537-4553.

[16] Li, Z., Zou, C., Gong, Z. and Wang, Z. (2014). The Computation of Average Run Length and Average Time to Signal: an Overview. *Journal of Statistical Computation and Simulation,* 84 (8), 1779-1802.

[17] Lucas, J. M. and Saccucci, M. S. (1990). Exponentially Weighted Moving Average Control Schemes Properties and Enhancements. *Technometrics,* 32, 1-29.

[18] Montgomery, D. C. (2013). *Introduction to Statistical Quality Control*, 7th ed. John Wiley & Sons: New York.

[19] Pickands, J. and Stine R. A. (1997). Estimation for an M/G/∞ Queue with Incomplete Information. *Biometrika,* 84, 295-308.

[20] Qiu, P. and Li, Z. (2011). On Nonparametric Statistical Process Control of Univariate Processes. *Technometrics,* 53(4), 390-405.

[21] Qiu, P., Zou, C. and Wang, Z. (2010). Nonparametric Profile Monitoring by Mixed Effects Modeling (with discussion). *Technometrics,* 52, 265-277.

[22] Reynolds, J. F. (1968). On the Autocorrelation and Spectral Functions of Queues. *Journal of Applied Probability,* 5, 467-475.

[23] Robert, S. W. (1959). Control Chart Test Based on Geometric Moving Averages. *Technometrics,* 1, 239-250.

[24] Ross, J., Taimre, T. and Pollett, P. (2007). Estimation for Queues from Queue Length Data. *Queueing Systems,* 55, 131-138.

[25] Shore, H. (2006). Control Charts for the Queue Length in a G/G/S System. *IIE Transactions,* 38, 1117-1130.

[26] Shu, L., Jiang, W. and Wu, Z. (2012). Exponentially Weighted Moving Average Control Charts for Monitoring Increases in Poisson Rate. *IIE Transactions,* 44, 711-723.

[27] Sonesson, C. and Bock, D. (2003). A Review and Discussion of Prospective Statistical Surveillance in Public Health. *Journal of the Royal Statistical Society, Series A,* 166, 5-21.

[28] Su, Y., Shu, L. and Tsui, K. L. (2011). Adaptive EWMA Procedures for Monitoring Processes Subject to Linear Drifts. *Computational Statistics and Data Analysis,* 55, 2819-2829.

[29] Tsung, F., Zhao, Y., Xiang, L. and Jiang, W. (2006). Improved Design of Proportional Integral Derivative Charts. *Journal of Quality Technology,* 38, 31-44.

[30] Zhou, Q., Zou, C., Wang, Z. and Jiang, W. (2012). Likelihood-Based EWMA Charts for Monitoring Poisson Count Data with Time-Varying Sample Sizes. *Journal of American Statistical Association,* 499, 1049-1062.

[31] Zi, X., Zou, C., Zhou, Q. and Wang, J. (2013) A Directional Multivariate Sign EWMA Control Chart. *Quality Technology and Quantitative Management,* 10, 115-132.

[32] Zou, C. and Tsung, F. (2010). Likelihood Ratio-Based Distribution-Free EWMA Control Charts. *Journal of Quality Technology,* 42, 174-196.

Authors' Biographies:

Dequan Qi is doctoral student of the Department of Statistics, School of Mathematical Sciences, Nankai University. His research interests include statistical process control.

Zhonghua Li is Assistant Professor of the Institute of Statistics, Nankai University. His research interests include statistical process control and quality engineering.

Xuemin Zi is Associate Professor of the School of Science, Tianjin University of Technology and Education. Her research interests include statistical process control and design of experiments.

Zhaojun Wang is Distinguished Professor and Vice Dean of Institute of Statistics, Nankai University. His primary research interests include statistical process control, quality improvement, and high-dimensional data analysis. His research has been published in various refereed journals including *Journal of the American Statistical Association, Technometrics, Journal of Quality Technology, IIE Transactions, Statistica Sinica, Naval Research Logistic*, etc.