DOI：10.13718/j.cnki.xdzk.2022.04.024

基于 Q 图的 $M(t)/M/C$ 服务台配置优化

高琦，　宫丽芳，　武艳华，　戴洪帅

山东财经大学 统计学院，济南 250000

摘要：排队问题随处可见，如何科学有效地对排队拥堵状况进行监控，合理配置资源具有重要的理论和现实价值. 与平稳模型相比，动态排队模型由于其时变特征，适用性更广. 本文根据逐段平稳和 $M/M/C$ 模型的相关理论，构建了基于 Q 图的动态 $M(t)/M/C$ 模型. 该模型利用 Q 控制图对过程中的到达率 λ 进行实时监测，根据变点对到达过程进行分段，进而优化服务台的配置问题. 随机模拟结果表明，该模型具有优良的检测效果，能够使有限的服务台资源得到有效利用. 最后，本文基于收费站数据，进行了实证分析，结果表明该模型可以实时监测到达率的变化情况，从而达到合理安排人工收费车道的目的.

关　键　词：Q 图；$M(t)/M/C$ 模型；逐段平稳

中图分类号：C931.1；O226　　**文献标志码**：A

文 章 编 号：1673－9868(2022)04－0206－07

开放科学(资源服务)标识码(OSID)：

Staffing Level for Dynamic $M(t)/M/C$: A Q-Chart Method

GAO Qi,　GONG Lifang,　WU Yanhua,　DAI Hongshuai

School of Statistics, Shandong University of Finance and Economics, Jinan 250000, China

Abstract: Queueing problems can be seen everywhere. How to effectively monitor queueing congestion and rationally allocate resources have important theoretical and practical values. The study of dynamic queues with time-varying arrival rates has become popular in queueing theory. In this work, we applied the Q-chart to study the staffing level for dynamic queueing system $M(t)/M/C$. This chart was applied to monitor the arrival rate λ in the process in real time. According to the change points, the process was segmentized to optimize the configuration of service station. Stochastic simulation results show that the method has advantage in improving the utilization rate of the service station. An empirical analysis was conducted in this study.

Key words: Q-Chart; $M(t)/M/C$ model; piecewise stationarity

收稿日期：2021－03－27

基金项目：国家自然科学基金项目(11901347)；山东自然科学基金项目(ZR2019MA035，ZR2020MA036).

作者简介：高琦，硕士研究生，主要从事排队系统的统计推断方面的研究.

通信作者：戴洪帅，教授.

排队系统的研究大体可分为 3 个方面：① 性能分析；② 统计推断；③ 资源配置[1]. 在排队模型的研究中，目前学者们通常假设顾客的到达率是不变的，具有平稳性. 但在很多情况下，如医院急诊室[2]，顾客的到达具有时变的特征，文献[3]指出若到达率只是偶尔非平稳，利用平稳排队模型分析系统的性能指标，得到的效果并不是最优的. 文献[4]指出即使到达率只是适度的不平稳，若利用平稳排队模型拟合系统，也会造成性能指标的严重低估，因此，逐段近似的方法成为研究非平稳排队模型的一类重要方法. 最近几十年，具有时变到达率的排队模型的资源配置问题成为学者们的研究热点之一. 文献[5]研究了非平稳到达过程的一般模型，使用平方根公式建立了服务台配置算法. 文献[6]根据改进的遗传算法建立了有关立体车库的排队模型，该模型可以缓解城市交通的压力，从而可以提升立体车库的服务率. 文献[7]为了实现资源利用效率的最大化，提出了以能量最小化为基础的虚拟机簇分配方法.

近年来，排队模型的统计控制问题引起了学者们的广泛关注. 文献[8]通过转移概率，给出变点监测的方法，并为 $M/G/1$ 和 $GI/M/1$ 模型系统利用率的监测提供了控制限. 文献[9]通过 CUSUM 表监控排队系统 $M/M/1$ 队列的服务性能，并与几种替代方法进行了对比. 考虑到数据的自相关性，文献[10]提出了一种基于加权似然比检验的 WLRT 控制图，该控制图能有效地监控排队系统的性能指标，尤其是对于 $M/M/1$ 队列系统利用率的监控.

目前在排队论的研究中，学者们通常把排队系统的性能监控和资源配置优化问题分开考虑. 另外，现实中很多排队系统的服务效率主要与服务台本身有关，如超市收银台的服务速率主要与收银员的熟练程度有关[11]. 基于以上考虑，本文将系统性能的监控和服务台配置综合考虑，基于 Q 图研究 $M(t)/M/C$ 模型服务台优化配置的问题.

1　理论基础

在 $M(t)/M/C$ 模型中，到达过程服从 Poisson 分布，但此时到达率不再恒定而是关于时间 t 的函数. 本文采用 $M/M/C$ 模型和 Q 图技术对动态 $M(t)/M/C$ 排队模型到达率的变化进行检测，从而优化服务台的配置问题，下面对符号术语进行简要的介绍.

1.1　$M/M/C$

本文采用逐段近似的方式优化 $M(t)/M/C$ 模型的服务台配置问题，通过监测变点对排队系统进行合理的分段，每段近似服从平稳的 $M/M/C$ 模型，并利用该模型研究相关问题. 在 $M/M/C$ 模型中，顾客的到达速率服从参数为 λ 的泊松分布，服务时间服从参数为 μ 的指数分布，服务台的数量为 C，等待空间和客源数量为无穷大，服务规则是先到先服务[12]. 其中，$\rho_c = \lambda/(C\mu)$ 表示系统的服务强度，一般来说只有当 $\rho_c < 1$ 时才能保证系统的平稳运行. W 是用来刻画顾客等待时间的参数，其平均值为 $E(W \mid W > 0)$，概率为 $P(W \leqslant T \mid W > 0)$. 进一步有

$$E(W \mid W > 0) = 1/[C\mu(1-\rho_c)] \tag{1}$$

$$P(W \leqslant T \mid W > 0) = 1 - \exp[-C\mu T(1-\rho_c)],\ T > 0 \tag{2}$$

1.2　Q 图

传统控制图通常只能监测服从标准正态分布的数据，对于服从泊松过程的数据其监测性能较差. 文献[13]将 Q 统计量应用于自适应移动加权平均控制图中，将参数值进行标准化处理，对常规控制方法进行了改进. 文献[14]也对 Q 图进行了性能的改进，在绘制泊松属性图的工作中采用了近似标准化的控制图，通过泊松分布进行变换，使数据近似标准化，并给出检测泊松参数的 Q 图技术. 因此，本文采用改进后的 Q 图来检测排队系统到达速率的变化.

2　模型

在 $M(t)/M/C$ 模型中，顾客的到达服从强度为 $\lambda(t)$ 的非齐次 Poisson 分布. 为了对队列的到达速率

进行实时监测，本文构建了基于 Q 控制图的监测方法，通过逐段平稳的 $M(t)/M/C$ 模型对其服务台配置问题进行研究.

2.1 Q 统计量

本文首先对到达率 $\lambda(t_i)$ 进行估计并检测变点，根据检测出的变点时刻将排队队列分为 $i(i=1,2,\cdots,m)$ 段，每段的时间记为 t_i，经过 n_i 次观测，顾客的到达速率为 $\lambda(t_i)$，到达速率随着 n_i 的变化记为 $n_i\lambda(t_i)$. 当每段中顾客的到达速率和到达人数均相同时，n_i 和 $\lambda(t_i)$ 为常数，即 $n_i=n$，$\lambda(t_i)=\lambda_0$，计算可得每段顾客人数的均值为 $n\lambda_0$，方差为 $n\lambda_0$，利用传统的 3σ 控制图构造上下控制限，则

$$UCL=n\lambda_0+3\sqrt{n\lambda_0} \qquad CL=n\lambda_0 \qquad LCL=n\lambda_0-3\sqrt{n\lambda_0}$$

每段所需观测的最小样本数如下所示：

$$n\geqslant\frac{-\ln(\alpha_L)}{\lambda} \qquad \alpha_L=P(x_i>UCL)=1-F([UCL];n_i\lambda_i) \tag{3}$$

当 $\lambda(t_i)$ 随时间 t_i 变化时，上下控制限的误差增大，控制图的灵敏度降低. 为了提高控制图的检测性能，本文将服从泊松分布的顾客数量的观测值 x_i 转化为可以在标准化正态 Q 图上绘制的值，X_i 是不同时段所有顾客数量的观测值 x_i 的和，即 $X_i=x_1+x_2+\cdots+x_i$，N_i 是整个系统的总观测次数，n_i 是该时段观测的总次数，即 $N_i=n_1+n_2+\cdots+n_i$. 将 x_1，x_2，$\cdots$，x_i 转化为 Q 统计量，通过下式给出相应的变换：

$$Q_i=\varphi^{-1}(\gamma_i)，\text{其中 } \gamma_i=B(x_i;X_i;n_i/N_i) \tag{4}$$

2.2 服务台配置优化算法

本文利用逐段平稳方法研究 $M(t)/M/C$ 模型的服务台配置优化问题. 利用 Q 图检测到达速率 λ 的变化，将发生变点的时刻记为 t_1，在 t_1 处对到达过程进行分段处理，并分别估计出每段的到达速率 λ_1 和 λ_2 的值. 若估计得到的 λ_2 和 λ_1 相等，则说明产生了误报，此时是一个伪检验，不需要采取任何措施；若得到的 λ_2 与 λ_1 相差较大，表明到达速率 λ 发生了变化，收到报警信号后，通过到达率 λ_2 与 λ_1 构造新的到达率 λ 的值，根据平稳模型对服务台的数量进行调整，从而优化系统中顾客的拥堵程度和等待时间.

具体算法如下：

① 利用 Q 图检测整个排队过程中到达率的变点，得到发生变点的时刻 t_1 和到达率的估计值 λ_1；

② 基于到达率 λ_1 和给定的延迟概率 α，利用 $M/M/C$ 模型的相关性质，计算服务台的数量 C，即

$$P(W\leqslant T\mid W>0)=1-\exp[-C\mu T(1-\rho_c)]\leqslant\alpha，\ T>0 \tag{5}$$

通过上式可知，服务台数量 C 满足

$$C\leqslant\lambda/\mu-\ln(1-\alpha)/(\mu T) \tag{6}$$

③ 在时刻 t_1 之后，基于到达率的估计值 λ_2，若 $\lambda_1=\lambda_2$，则将检测下一个变点，若 $\lambda_1\neq\lambda_2$，则重复步骤 ② 计算该段服务台的最优配置；

④ 重复上面的步骤，直至周期结束.

3 模拟实验

本节设计了随机模拟实验证明该模型的理论优势. 首先选取 i 个样本观测值，对 λ 设定不同的参数来模拟 Q 图的监测过程. 本文共选取了48个样本点，即 $i=1,2,\cdots,48$. 根据式(3)计算得到样本的最小值为4，因此，可将前40个样本观测值每个节点的观测次数 $n_i=4$，后8个样本观测值的每个节点的观测次数 $n_i=5$. 到达速率初始值设为 $\lambda_1=1.7$，$\lambda_2=3.4$，服务速率设为 $\mu_1=\mu_2=1$. 假定 $i=39$ 时，到达速率 λ 发生变化，即前39个观测点服从 $\lambda_1=1.7$ 的泊松分布，后9个观测点服从 $\lambda_2=3.4$ 的泊松分布. 模拟得到系统中每个节点的观测次数 n_i 和人数 x_i 的取值如表1所示.

表 1　模拟数据

n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i
4	5	4	11	4	8	4	6	4	5	4	7	4	9	5	8
4	6	4	8	4	6	4	9	4	6	4	7	4	4	5	19
4	7	4	8	4	9	4	11	4	2	4	7	4	15	5	21
4	10	4	3	4	7	4	5	4	6	4	4	4	16	5	14
4	5	4	5	4	8	4	8	4	10	4	9	5	20	5	15
4	10	4	4	4	14	4	4	4	6	4	8	5	17	5	18

经计算可得

$$X_i = x_1 + x_2 + \cdots + x_{48} = 420 \qquad N_i = n_1 + n_2 + \cdots + n_{48} = 200$$

通过式(4)，计算得到的 Q 分位数结果如表 2 所示.

表 2　Q 分位数

i	Q_i	i	Q_i	i	Q_i	i	Q_i	i	Q_i	i	Q_i	i	Q_i	i	Q_i
1	−0.705	7	−0.045	13	0.782	19	1.246	25	−2.021	31	0.141	37	2.681	43	3.018
2	0.484	8	−1.854	14	0.068	20	−0.849	26	−0.232	32	−1.024	38	2.871	44	1.155
3	−1.116	9	−0.857	15	0.424	21	0.285	27	1.166	33	0.873	39	3.822	45	1.394
4	0.743	10	−1.158	16	2.237	22	−1.211	28	−0.250	34	0.514	40	2.174	46	2.126
5	0.921	11	0.426	17	−0.427	23	−0.750	29	0.134	35	0.845	41	−0.469	47	NA
6	−0.086	12	−0.289	18	0.643	24	−0.329	30	0.138	36	−1.034	42	2.624	48	NA

在 Q 控制图中描绘出各模拟点的值，从图 1 可以看出不同点对应的 Q 分位数的波动. 其中，第 39 个点对应的分位数超出了控制限[−3，3]的范围，系统发出警报，表明该控制图检测出了 λ 的变化.

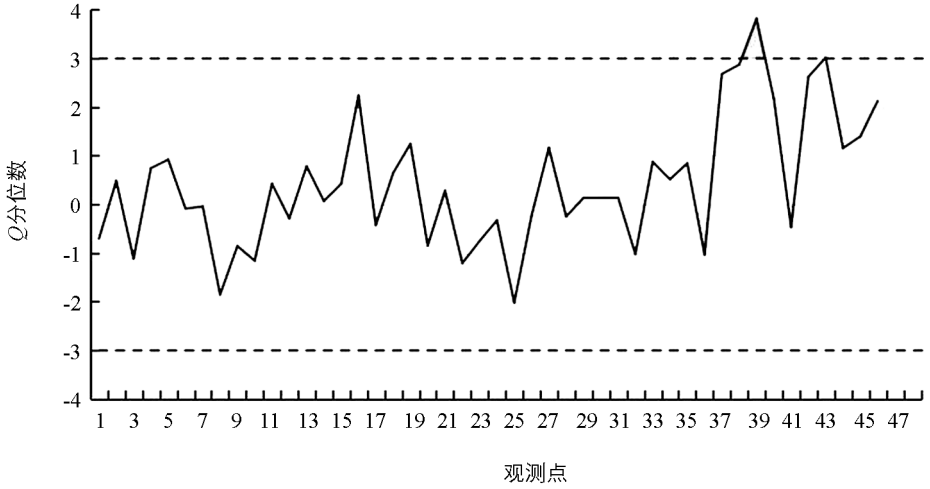


图 1　模拟数据的 Q 控制图

由于式(4)中的二项分布是泊松分布的最小方差无偏估计[14]，因此可以用二项分布的概率 $p_i = n_i/N_i$ 来估计到达速率 λ_i，得到 $\lambda_1 = 1.756\ 998$，$\lambda_2 = 3.358\ 514$. 其中，估计得到的 λ_1 和 λ_2 相差较大，说明没有发生误报. 第 39 个样本点是发生变化的点，根据逐段平稳的性质对系统进行分段处理，将前 38 个样本观测点规划到前一个节点内，后 10 个样本观测点规划到后一个节点内，得到权重 ω_1 和 ω_2 的值.

因此，根据优化后模型的权重计算到达速率 λ 的值为

$$\lambda = \omega_1\lambda_1 + \omega_2\lambda_2 = 2.090\ 647$$

通过算术平均计算得到的到达率 λ' 的值为

$$\lambda' = (\lambda_1 + \lambda_2)/2 = 2.557\ 756$$

基于Green等标准的服务水平优化目标[15]，本文设置延迟概率不超过80%，即$\alpha = 0.8$，可以得到等待时间小于T的概率，

$$P(W \leqslant T \mid W > 0) \leqslant 0.8$$

为了使顾客的等待时间相对较少，令$T = 10/60 = 1/6$，由此计算出满足此条件的服务台数量为$C \leqslant 11.74$，这里服务台个数取整为12. 由式(1)可以得出排队系统中每一位顾客的平均等待时间为$W = 0.100\ 9$. 同理根据λ'计算得出其服务台数量和等待时间分别为$C' \leqslant 12.21$，$W' = 0.105\ 9$，服务台个数取整为12. 为了能够合理地分配总的服务台个数，根据系统的逐段平稳性质得到第一个节点和第二个节点所需要的服务台数量为

$$C_1 = \frac{\lambda_1}{\mu} - \frac{\ln(1-\alpha)}{\mu T} = 11.41,\ C_2 = \frac{\lambda_2}{\mu} - \frac{\ln(1-\alpha)}{\mu T} = 13.02$$

通过模拟，根据C_1和C_2的取值，可以更加合理地安排不同时段的服务台数量，第一段安排服务台数量为11，等待时间$W_1 = 0.097\ 6$，第二段安排服务台数量为13，等待时间$W_2 = 0.093\ 9$.

2种模型计算的结果如表3所示.

表3 模拟结果对比表

	λ	C	ρ_C	W
算术平均	2.558	12	0.213	0.105 9
节点一	1.757	11	0.146	0.097 6
节点二	3.359	13	0.240	0.093 9

对比这2种模型所需服务台的数量和顾客的等待时间，研究发现优化之后的模型更加实用，主要体现在以下2个方面：

① 从顾客角度来说，优化之后的模型可以减少每一位顾客的等待时间，进而可以提高顾客的满意度.

② 从资源利用的角度来说，优化后的模型能够根据顾客到达率的变化合理安排服务台的个数，可以更为有效地利用服务资源，完成资源的优化配置.

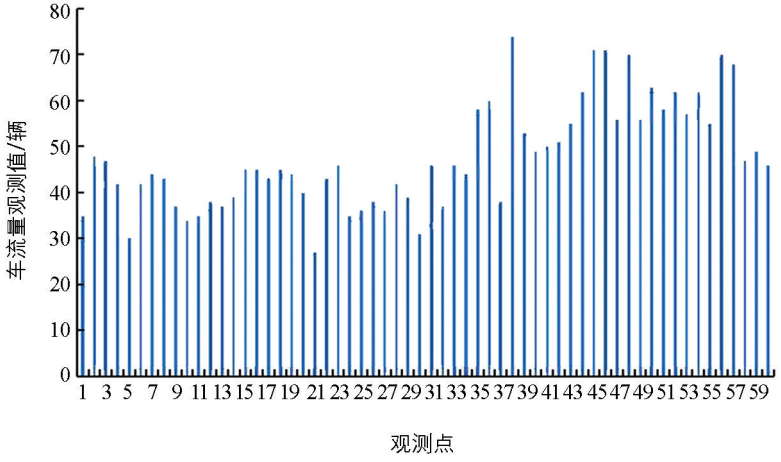
4 实证分析

本文选取与现实生活密切相关的高速公路人工收费车道问题进行实证分析. 随着车辆的不断增加，高速公路收费站处会出现到达车辆不能及时得到服务的状况，为了解决车辆排队的拥堵现象，文献[16-17]分别采用$M/M/N$和$M/G/K$模型系统描述收费站的排队情况. 本文以山东省某收费站为例，利用上述模型对车流量进行监测，并合理安排人工收费车道的数量. 该收费站共有8个服务台，3条ETC通道，5条人工通道，由于ETC通道全天候开放，故只对人工通道的车流量进行观测并记录，其中工作人员实行三班轮换制，即每8 h换班一次. 由于每周的车流量数据具有一定的周期性，因此每周只选取一天进行观测，连续观测6周，于每周三6：00－21：00每15 min记录一次观测值，发现同一时段的观测值相差不大，随机选取收费站某天的车流量数据进行实证分析.

各时段车流量数据共计60个观测值，其分布直方图如图2所示. 为了对各时段车流量数据的分布进行研究，采用K-S检验法，由图2可知P值为0.058，因此该数据近似服从泊松分布. 其中前40个观测值的$n_i = 4$，后20个观测值的$n_i = 5$，$\mu \approx 50$辆/h，$T = 1/6$ h，车流量的Q图如图3所示.

利用Q图对数据进行监测，图3表明当第1、第3和第36个点超出控制限时，系统发生报警. 由于开始时系统的不平稳特征[18]，我们排除第1和第3个点的干扰，即无需改变人工收费站的数量. 因此以第36个点作为变点进行分段处理，由极大似然估计可得$\lambda_1 = 165$辆/h，$\lambda_2 = 233$辆/h，经计算可得，$C = 5$，$W \approx$

0.018 h，$\lambda=193$ 辆 /h，$\lambda'=199$ 辆 /h，$C_1=4$，$C_2=5$. 该模型的计算结果表明，在变点之前安排 4 个人工收费车道，在变点之后安排 5 个人工收费车道，相比较于收费站全天候开放 5 个人工收费车道，优化之后的模型可以及时调整人工收费车道的开放个数，从而能够降低人工费用等各项支出.



单样本Kolmogorov-Smirnov(K-S)检验

		VAR00001
数字		60
泊松参数 a,b	平均值	47.833 3
最极端差分	绝对	0.172
	正	0.172
	负	0.139
Kolmogorov-Smirnov Z		1.330
渐进显著性（双尾）		0.058

a. 检验分布是泊松分布

b. 根据数据计算

图 2　车流量分布与 K-S 检验图

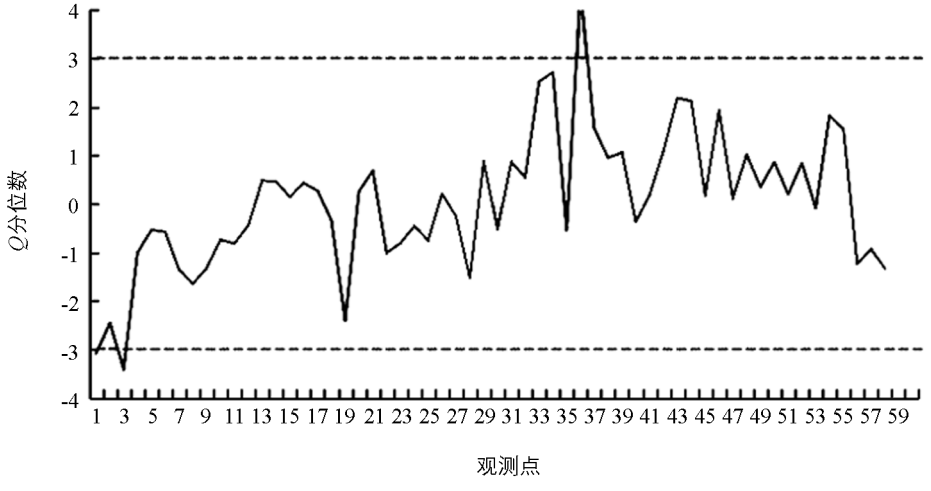


图 3　车流量的 Q 控制图

5　结论

本文基于 Q 控制图对到达速率 λ 的动态变化进行监控，通过控制图找到异常点，根据变点时刻分段，得到逐段平稳的排队列，利用 $M(t)/M/C$ 模型的相关理论知识计算出服务台的数量和等待时间，合理地安排服务台数量. 此方法的优点是将变点检测和逐段平稳相结合，从而优化了动态排队模型的服务台配置问题. 本文利用收费站某天的车流量数据进行了实证分析，结果表明，Q 控制图能够对动态的车流量数据进行实时监测，检测出车流量变化幅度较大的点. 根据变点计算出各段的人工收费车道数量，进而实时调整服务资源，缓解收费站的拥堵情况，降低收费站人工成本等各类的支出，具有重要的现实意义.

参考文献：

[1] 唐加山. 排队论及其应用 [M]. 北京：科学出版社，2016.

[2] LAKSHMI C，LYER S A. Application of Queueing Theory in Health Care：a Literature Review [J]. Operations Research for Health Care，2013，2(1-2)：25-39.

[3] LIU Y N. Many-Server Queues with Time-Varying Arrivals, Customer Abandonment and Non-Exponential Distributions [D]. New York: Columbia University, 2011.

[4] GREEN L, KOLESAR P. The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals [J]. Management Science, 1991, 37(1): 84-97.

[5] HE B X, LIU Y N, WHITT W. Staffing a Service System with non-Poisson non-Stationary Arrivals [J]. Probability in the Engineering and Informational Sciences, 2016, 30(4): 593-621.

[6] 李建国, 张海飞, 周璐婕, 等. 基于改进遗传算法的立体车库布局对比及服务资源优化 [J]. 西南大学学报(自然科学版), 2019, 41(4): 139-148.

[7] 邱春红. 云计算中虚拟机簇的优化分配 [J]. 西南师范大学学报(自然科学版), 2021, 46(1): 44-49.

[8] BHAT U N, RAO S S. A Statistical Technique for the Control of Traffic Intensity in the Queuing Systems M/G/1 and GI/M/1 [J]. Operations Research, 1972, 20(5): 955-966.

[9] CHEN N, ZHOU S Y. CUSUM Statistical Monitoring of M/M/1 Queues and Extensions [J]. Technometrics, 2015, 57(2): 245-256.

[10] QI D Q, LI Z H, ZI X M, et al. Weighted Likelihood Ratio Chart for Statistical Monitoring of Queueing Systems [J]. Quality Technology & Quantitative Management, 2017, 14(1): 19-30.

[11] 王颖俐. 基于M/M/c/∞排队模型分析超市收银台数量 [J]. 太原师范学院学报(自然科学版), 2015, 14(2): 8-10, 27.

[12] SYSKI R. Fundamentals of Queueing Theory [J]. Technometrics, 1999, 41(1): 76-77.

[13] 权政, 赵玲玲, 徐滨, 等. 面向大规模定制的改进型AEWMAQ控制图应用研究 [J]. 机床与液压, 2021, 49(24): 8-12.

[14] QUESENBERRY C P. SPC Q Charts for a Binomial Parameter P: Short or Long Runs [J]. Journal of Quality Technology, 1991, 23(3): 239-246.

[15] GREEN L V, KOLESAR P J, WHITT W. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System [J]. Production and Operations Management, 2007, 16(1): 13-39.

[16] 姬杨蓓蓓, 周金凤. 基于成本分析的高速公路收费站车道配置研究 [J]. 重庆交通大学学报(自然科学版), 2018, 37(1): 85-91.

[17] 曲明革. 高速公路出入口收费车道数研究 [J]. 公路, 2012, 57(5): 262-267.

[18] 黄虎, 柯华, 王晶. 基于逆加权参数估计方法的改进型Q控制图研究 [J]. 系统工程学报, 2016, 31(4): 568-574.

责任编辑 崔玉洁