Conference Paper

Research on Library Queuing Model Based on Data Mining

Longjie Sun and Kaijun Yu

Library of Shanghai University of Medicine & Health Sciences, 279 ZhouZhu Road, Shanghai, China

Abstract

The current work of library circulation service desk is being transferred from traditional pure manual operations to human–machine collaboration and artificial intelligence. It is urgent to study a set of general and reliable service desk distribution models to better optimize staffing and improve service efficiency and quality. Based on the theory of queuing, the model for the optimization of distribution desks is explored, and the overall design of the plan is completed (sample data collection and mining, parameter estimation, operation index calculation, results analysis, and evaluation, etc.). In this case, the parameters were selected from Nanyuan library circulation service desk in Shanghai University of Medicine & Health Sciences. Based on actual statistical sampling, this article has estimated the feasibility, reliability, and effectiveness by a reasonable parameter-range verification scheme, provided a strong reference for the decision-making of library circulation departments, and effectively improved the efficiency of circulation services.

Keywords: queuing theory, data mining, stratified sampling

Corresponding Author:
Longjie Sun
sunlj@sumhs.edu.cn

Received: 29 August 2018
Accepted: 18 September 2018
Published: 11 November 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICOI-2018 Conference Committee.

1. Introduction

With the rise of artificial intelligence and the change of human resources in university libraries, the traditional circulation service has changed dramatically in recent years[1-4], traditional paper books and periodic loan repayments show a trend of decreasing, the content of the service is more diversified and intelligent, and the queuing is more humanization. These new requirements for circulation services are also a new topic for librarians. The research on circulation queuing at home and abroad has considerable depth, mainly focusing on business, logistics, transportation, society, medical treatment[5-9]. Queuing theory is also widely studied in the field of library, such as the best copy quantity, technology investigation, personnel scheduling and so on [10]. Current research on distribution services has also made positive progress[11-13], such as basically agree with the system of multi-circulation service desk in library should be the M/M/C model of queuing theory, determine the arrival rate of key parameters λ, service rate μ corresponding to the number of readers arriving at the service desk

per unit of time and the number of service units per service station per unit time and a series of operating index parameters[14-15]. However, the actual estimation of the parameter arrival rate λ and service rate μ still lacks effective solutions.

This paper proposes research objectives and specific solutions for the above issues. Research objectives: Determine the number of best service desks that readers do not backlog in queue during peak hours, or the queuing probability is in an acceptable range; Determine the number of best service desks that employees are not free during normal hours, or the probability of idleness is controlled within a certain range. The solution is as follows.

1. Determine the applicable queuing model based on the actual number of service desks and personnel arrival characteristics;
2. Using data mining [16, 17], sampling [18] and other technologies to process the flow data, giving the estimation process and recommended values of the main parameters λ, μ;
3. Calculate the parameters of major operational indicators, analyze and determine the best number of service desk, and analyze the actual application value of the scheme through the comparison between theory and practice;
4. To sum up, further suggestions on high operability, high applicability and some issues that need attention have been put forward.

2. Program flow

In the case of the multi-service desk queuing model (i.e., M/M/C), the specific calculation steps are shown in Figure 1 below.

2.1. Data mining

The estimated values of the main parameters λ, μ can be randomly sampled from the existing library access control system and OPAC query system. The data exported from the background is cluttered, repetitive, or incomplete because of card delay, self-sensitivity of the access control, and card reader settings, therefore, data mining and collation must be performed on the sample data before random sampling to improve the quality of the collected data and estimation reliability.

We imported the data of the internal access flow data (excel sheet) from the initial period (240 days record) into the SPSS Model for data modeling. The ascending or

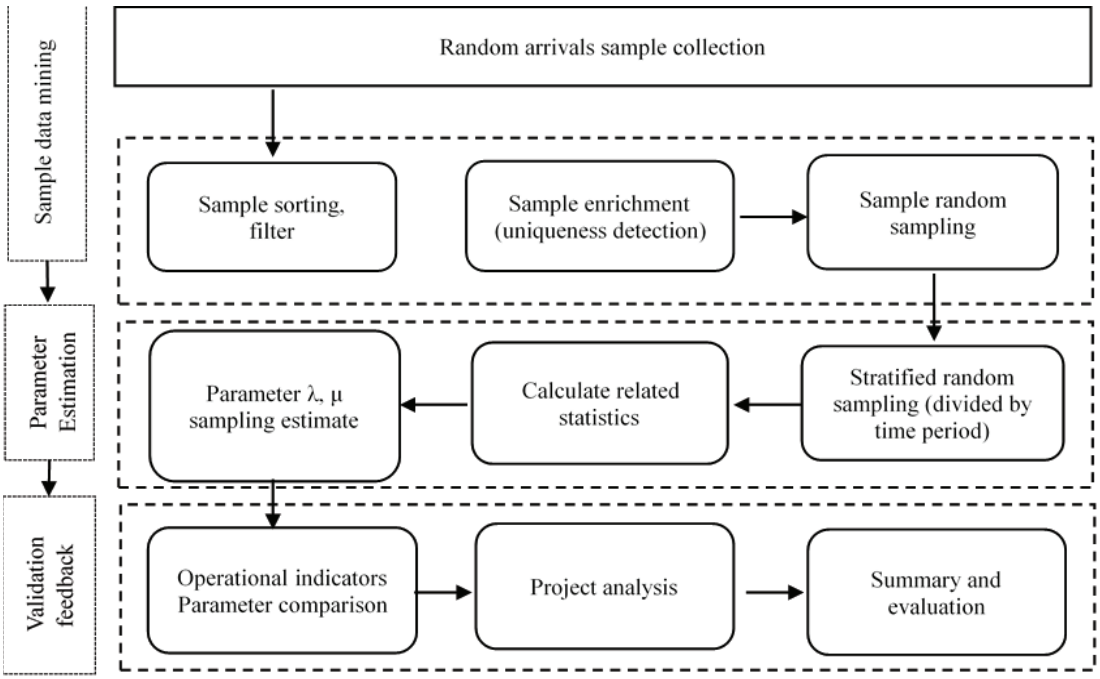

Figure 1: Flow chart.

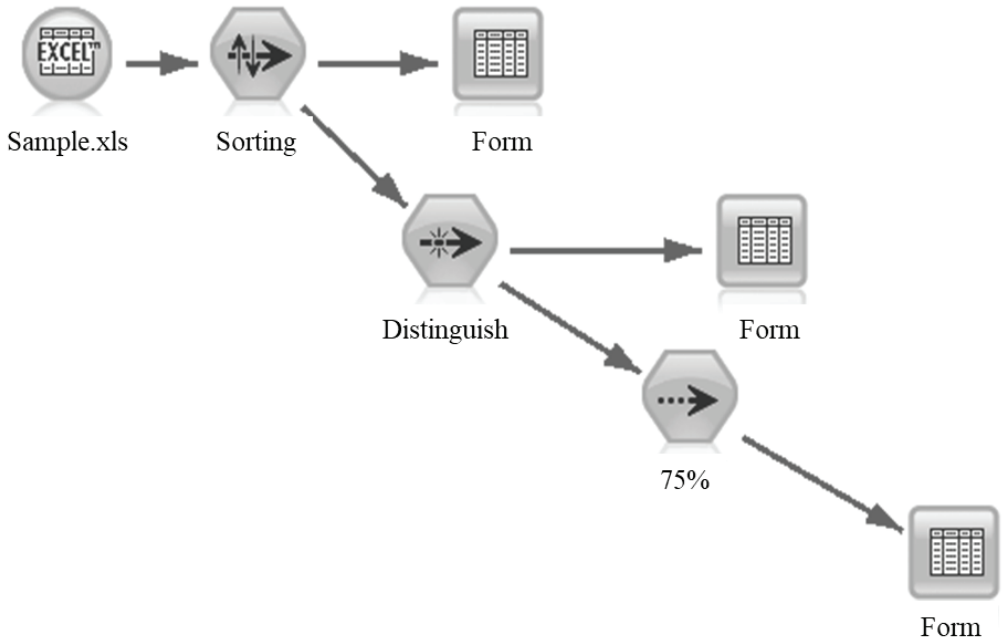

Figure 2: Data Mining.

descending order of the sample data can be achieved by setting the time parameter

in the Sort node in Record Ops. The sample selection is implemented through the Select node in Record Ops. The actual extracted data in this scheme contains more redundant data and must be checked for repeatability, this can be achieved through the parameter setting of the Distinct node in Record Ops. After obtaining the circulation sample for the period according to the above data mining, a certain percentage of samples are sampled according to the random principle by the Sample node in Record Ops. We select 75% (i.e. 180 days) as a random sample overall size N. The specific process is shown in Figure 2.

2.2. Parameter estimation

2.2.1. Stratified random sampling

Through long-term observation, the library readers' arrival rules are not completely random, but are greatly affected by the course arrangement. Therefore, it is necessary to stratify several representative levels at the open daily time.

Here is a simple explanation of the related properties theorem of stratified sampling. In order to facilitate discussion, the relevant symbols of stratified sampling need to be defined. The estimates of all population parameters are marked with the "st" subscript. st is the shorthand of "stratified", meaning stratification. See table 1 for details.

Table 1: Related Symbol Description.

Symbol	h	i	N_h	n_h	Y_{h_i}	y_{hi}
Meaning	Subscript "Level h"	Subscript "unit number"	Total number of units in layer h	Number of samples in layer h	The value of the i-th overall cell in layer h	The i-th sample cell value in layer h
Symbol	w_h	f_h	$\bar{Y}_h$		$\bar{y}_h$	
Meaning	$\frac{N_h}{N}$ layer rights of Level h	$\frac{n_n}{N_h}$ Sampling ratio of Level h	$\frac{1}{N_h}\sum_{i=1}^{N_4} Y_{h_i}$ Overall average of Level h		$\frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_j}$ Sample mean of of Level h	
Symbol	Y_h	y_h	S_h^2		s_h^2	
Meaning	$\sum_{i=1}^{N_h} Y_{h_i} = N_h\bar{Y}_h$ Total amount of Level h	$\sum_{i=1}^{n_h} y_{hi} = n_h\bar{y}_h$ Total samples number of Level h	$\frac{\sum_{i=1}^{Nh}(Y_{hi}-\bar{Y}_h)^2}{N_h-1}$ The overall variance of Level h		$\frac{\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2}{n_h-1}$ Sample variance of Level h	

Stratified sampling first calculates an appropriate estimate of the average value of each layer according to the sample of each layer. Then by the layer estimate , the weighted average of the overall layer weight is estimated by, i.e.,

$$_{st} = \sum_{h=1}^{L} W_h = \frac{1}{N} \sum_{h=1}^{L} N_h \tag{1}$$

For stratified random sampling, the samples in each layer are independently performed according to simple random sampling, is taken as the sample mean of the h-layer, and the simple estimate of is recorded as $\overline{\overline{\sum}}$.

$$\hat{\overline{Y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} N_h \tag{2}$$

For stratified random sampling, $\overline{\overline{\sum}}$ is an unbiased estimate of . For stratified random sampling, the unbiased estimate of variance $\bar{v}(\bar{y}_{st})$ for $\overline{\overline{\sum}}$ is,

$$\bar{v}(\bar{y}_{st}) = \sum_{h=1}^{L} \mathrm{W}_h^2 \frac{1-f_h}{n_h} \mathrm{s}_h^2 = \sum_{h=1}^{L} (\frac{1}{n_h} - \frac{1}{N_h}) \mathrm{W}_h^2 \mathrm{s}_h^2 = \sum_{h=1}^{L} \frac{\mathrm{W}_h^2 \mathrm{s}_h^2}{n_h} - \sum_{h=1}^{L} \frac{\mathrm{W}_h \mathrm{s}_h^2}{\mathrm{N}}$$

In the formula, $\mathrm{s}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y})$ is the sample variance of the h-th layer sample. For stratified random sampling, the simple estimate of total population Y is $\hat{Y}_{st} = \mathrm{N}\bar{y}_{st}$, which has the following properties:

$$E(\hat{Y}_{st}) = Y \tag{3}$$

$$V(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h() \tag{4}$$

$$\hat{v}(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h() \tag{5}$$

That is, $\hat{v}(\hat{Y}_{st})$ is an unbiased estimate of $\mathrm{V}(\hat{Y}_{st})$.

2.2.2. λ parameter estimation

The total sample data (180 days) obtained from data mining is divided into five levels by period, each level is sampled independently, and each level is a simple random sampling, the number of people arriving within 15 days is counted. Specifically shown in the table 2 below. According to the sampling formulas and the sampling data in table

Table 2: Random sampling of arrivals.

Level(h)	Randomly selected for 15 days														
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1(8 : 00-9 : 30)	30	40	29	22	38	36	18	51	36	25	11	14	19	20	47
2(9 : 30-11:30)	55	44	63	56	95	22	18	107	89	108	43	114	23	28	45
3(11 : 30-13 : 00)	120	113	89	99	142	75	143	170	356	165	131	119	53	104	41
4(13 : 00-14 : 30)	136	192	200	84	161	85	58	205	276	234	204	320	104	51	111
5(14 : 30-16 : 30)	105	139	96	63	146	45	36	102	100	134	130	89	17	13	33

Table 3: The describes statistics of arrivalS.

h	N_h	n_h	f_h	w_h	Min	Max	$\bar{y}_h$	$N_h\bar{y}_h$	S_h^2
h1	180	15	0.083	0.2	11	51	29.07	5232.6	144.638
h2	180	15	0.083	0.2	18	114	60.67	10920.6	1136.381
h3	180	15	0.083	0.2	75	356	151.47	27264.6	5234.267
h4	180	15	0.083	0.2	51	320	161.40	29052	6491.971
h5	180	15	0.083	0.2	13	146	83.20	14976	2075.886
Total	900	75	——	1.00	——	——	——	87445.8	——

2, combined with statistical analysis of SPSS software, statistical statistics are obtained, as shown in Table 3.

is the estimate of the population mean in 180 days, can be obtained by formula 1, $\overline{Y}_{st} = \sum_{h=1}^{5} N_h$ = 87445.8.λ needs further solution, here we calculate per capita statistics for per day (8.5 hours), $\overline{\lambda}_1 \frac{87445.8}{180\times8.5} \approx$.

In order to further improve the reliability of estimation, it is necessary to estimate the 95% confidence interval range of λ, where $z_{0.025} = 1$ is given by formula 6.

$$\hat{v}() = \sum_{h=1}^{5} N_h() \approx 29864623.1 \tag{6}$$

$$5464.85 \tag{7}$$

The 95% confidence interval for is,

$$76734.68 \tag{8}$$

$$\overline{Y}_{st} - z_{0.025}\sqrt{\hat{v}(\hat{Y}_{st})}98156.912 \tag{9}$$

Range of is 50~64.

According to the number of statistics in table 3, there is a large amount of data in the peak period of university library, which needs to be extracted separately for reference, this example is clearly concentrated in the third and fourth layers (11:30-14:30).

$$= 4818.5 \tag{10}$$

The 95% confidence interval for is,

$$\hat{Y}_{st} - z_{0.025}\sqrt{\hat{v}(\hat{Y}_{st})} = 46872 \tag{11}$$

$$\hat{Y}_{st} - z_{0.025} \tag{12}$$

Range of is 87~122.

2.2.3. μ parameter estimation

According to the main types of daily service of circulation tables, they are divided into 3 layers. The sampling units are sampled according to a certain weighted proportion. Each layer is sampled by simple random sampling for 10 days. SPSS software is used to derive the relevant statistics, and 95% of μ is estimated.

Table 4: Service time random sampling.

level（h）	Sample	Randomly selected 10 days									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1(Generally borrowed)	180	19	9	15	12	9	10	8	7	10	7
2(Extended, postponed)	180	11	44	19	16	26	30	45	52	41	19
3(Add, modify, cancel)	40	65	110	120	49	40	27	22	29	56	44

Table 5: Service time statistics.

h	N_h	n_h	f_h	w_h	Min	Max	$\bar{y}_h$	$N_h\bar{y}_h$	S_h^2
h1	180	10	0.056	0.45	7	19	10.60	1908	14.489
h2	180	10	0.056	0.45	11	52	30.3	5454	204.456
h3	40	10	0.25	0.1	22	120	56.2	2248	1140.844
Total	400	30	——	1.00	——	——	——	9610	——

According to the definition of μ, we need to calculate the number of people served in unit time, here we take the unit time as hours. In one hour, μ≈150.

$$\hat{v}(\hat{Y}_{st}) = \sum_{h=1}^{3} N_h(N_h - n) \tag{13}$$

$$\sqrt{\hat{v}(\hat{Y})} = 898.26 \tag{14}$$

The 95% confidence interval for is,

$$\hat{Y}_{st} - z_{0.025}\sqrt{\hat{v}(\hat{Y}_{st})} = 7849.41 \tag{15}$$

$$\hat{Y}_{st} - z_{0.025}\sqrt{\hat{v}(\hat{Y}_{st})} = 11370.5896 \tag{16}$$

Average service time range is $19.6 \sim 28, \mu \approx 184\ 127$.

2.3. Validation feedback

Based on the results and ranges of λ and μ obtained from stratified sampling, let c be the number of service desks, the corresponding service strength , idle probability Po and queuing probability P are calculated by the formula of queuing theory [19-20]. The specific formula is as follows.

$$(\rho < 1,\ \text{the } s) \tag{17}$$

$$P_0 = [] \tag{18}$$

$$= \tag{19}$$

$$(> c) = 1 - \sum_{h=0}^{c} P_n \tag{20}$$

Because the probability of service is relatively stable, μ is appropriate to take the median amount of 150. Table 6 shows the influence of the number of different service desks c and the number of arrivals λ on the idle probability and queuing probability under this parameter.

From the above table, it can be seen that the choice of the number of service desks generally depends on two types of factors: 1. The probability of idleness is controlled within a reasonable range; 2. The queues or the probability is not as low as possible during the peak period. If the idle probability is in the range of 30%~70%, the queuing

Table 6: System Operation Reference Table.

C / λ	1	2	3	4
57	P_0=62% P(>1)=14.44%	P_0=68.07% P(>2)=1.15%	P_0=68.37% P(>3)=0.08%	P_0=68.38% P(>4)=0.02%
104	P_0=30.67% P(>1)=48.07%	P_0=41.17% P(>2)=10.50%	P_0=49.85% P(>3)=0.84%	P_0=49.98% P(>4)=0.10%
160	ρ>1	P_0=30.43% P(>2)=19.78%	P_0=33.89% P(>3)=3.79%	P_0=34.33% P(>4)=0.68%

probability is in the range of 0~10% as the reference standard. When c=1, the queuing system is prone to instability during the peak period, resulting in a decline in the satisfaction of the circulation service and inconsistent with the purpose of "service and education". When c=2, 3, and 4, the indicators are more consistent. Longitudinal comparisons found that when c=2, the queuing probabilities at the peak time exceeded the set criteria, and when c=4, the queuing probabilities were small and there was no obvious difference, which was easy to waste resources. Therefore, the number of service desks c=3 is more appropriate.

Through the above-mentioned series of operation practices, the number of the best service desks in this paper is obtained, and the feasibility of the scheme is verified. The trial run after a certain period also basically complies with the probabilistic statistical results obtained, indicating that the scheme is reliable.

3. Conclusion and Evaluation

Although the above case is a case, but in view of various university libraries can be basically consistent operating rules. This case starts with the original access data and unceasingly excavates the related parameter estimation samples, according to sampling statistics and general formula, the actual parameters are obtained, calculates the number of service stations that meet the requirements, provides sufficient basis for library circulation management decision and personnel allocation.

The focus of this paper is how to obtain valuable key parameters through the mining of existing data, makes a solid foundation for the study of library circulation queuing models, provides a series of practical and practical methods and steps, makes the source of main parameters more scientific and reasonable. Due to the differences between the opening hours and the actual circulation, this model is only for reference, the peak judgment and the collection of service time need to be further improved.

References

[1] J. Zhou, Paper-based literature interviews based on (S,S) strategy - loan optimization analysis, Library and Information Service, 2011, 55(1): 60-64.

[2] Y. Li, The status quo, Literature borrowing status, problems and countermeasures in Fuzhou University City Library Union, Library Science Research, 2013, (6): 77-80

[3] D. Li, Y. Dong, L. Xie, An analysis of factors affecting the collection of paper resources in library based on user behavior, Information Science, 2014(7):103-107

[4] Pancheva, D., Mukhtarov, P., & Penov, N., PITFALLS IN STUDYING PLANETARY WAVES IN THE MIDDLE ATMOSPHERE DURING NORTHERN EXTRATROPICAL WINTER. *Comptes rendus de l'Académie bulgare des Sciences*, 2016, *69*(5).

[5] W.E. Biles, Design of simulation experiments, Proceedings of the 16th conference on winter simulation, Piscataway: IEEE Press,1984:98-104

[6] T. Smith, Queueing for a chance to live again, 2017(05):30.

[7] R.W. Wolff, Stochastic modeling and the theory of queues, Upper Saddle River: Prentice Hall,2000.

[8] H. Zheng, X. He, Research on traffic flow at intersections based on queuing theory, Science,2010(35):377-378.

[9] Y. Wu, S. Wang, C. Wang, Queuing theory in the application of logistics planning, Logistics Technology, 2004 (5): 53-55.

[10] Y. Xiong, S. Hu, The application of queuing theory in electronic document service system, Modern Library Information Technology, 2008(11):82-85.

[11] Q. Zhuge, J. Yao, The application of queuing theory in the library circulation department, Shanghai University Library Information Research, 2011(2):50-52.

[12] K. Cai, J. Guo, N. Zhai, Queuing theory analysis and solution of circulating circulation in university libraries, Science and Technology Information Development and Economy, 2014(23):6-8.

[13] C. Jiang, X. Yuan, L. Liu, University library circulation service desk configuration model, Library and Information Service, 2017, 61(20): 97-104.

[14] C. Lu, Queuing Theory, Beijing: Beijing University of Posts and Telecommunications Press, 1994

[15] Alexandrov, A. S., Vassileva, P., Momchilova, A., Tsonchev, Z., Kirilova, Y., Ivanova, R.,... & Orozova, M., A new approach using nanomembrane-based therapeutic plasmapheresis for treatment of patients with multiple sclerosis and neuromyelitis optica. *Comptes rendus de l'Académie bulgare des Sciences*, 2016, *69*(3).

[16] W. Xue, SPSS Modeler Data Mining Method and Application, Beijing: Publishing House of Electronics Industry, 2014.

[17] Yossifova, M., Dimitrova, D., & Iliev, T., Phase composition of dry residues from water leachates of coal, claystone partings and combustion wastes, maritsa east lignite basin, bulgaria. *Comptes rendus de l'Académie bulgare des Sciences*, 2016, *69*(12).

[18] Y. Jin, Z. Du, Y. Jiang, Sampling Technology, Beijing: China Renmin University Press, 2015.

[19] X. Li, Queuing Theory, Beijing University of Posts and Telecommunications Press, 1994

[20] Z. Han, L. Yu, Mathematical modeling method and its application, Beijing: Higher Education Press, 2010.