Optimization of Queueing Model for Local Bubble Tea Shop

Abstract — **Kung Fu Tea, located in Charlottesville, VA is a popular location for UVA students to order boba tea. Currently, the busiest shifts – 2:00pm to 6:00pm on the weekends – have 1 employee at the register and 3 drink makers, and our paper aims to model this as a Markov Chain with the queue being represented as a birth-and-death process. We proposed a new model, with 2 employees at the register and 2 drink makers, and found this to improve waiting times by 0.49 minutes on average. Therefore, we recommend that Kung Fu Tea adopt a new model of 2 employees at the register and 2 making drinks from 2:00pm to 6:00pm on the weekends.**

I. INTRODUCTION

Kung Fu Tea, located on 1001 W Main St, Charlottesville, VA 22903, is a popular boba tea destination for students at the University of Virginia. The undeniable popularity of boba paired with the convenience of the store means that drinks are often in high demand, putting pressure on the servers - who are often college students themselves - to make quality boba tea quickly and efficiently. The goal of this project is to implement an effective model to reduce waiting time and balking rate for customers while preserving service quality.

Quality and service are often two interfering factors when it comes to a business and restaurant establishments. The article "Exploring the role of service quality, atmosphere and food for revisits in restaurants by using a e-mystery guest approach"[2] explores these intersecting spheres by modeling them as an inverse equation. The study also goes to show that an employee's reliability has a direct impact on the atmosphere of the establishment as well as a customer's intention to revisit. In particular, the study aims to use the concept of "mystery guests" that evaluate the processes of the establishment while keeping the establishment in the dark about their presence. This ensures an accurate representation of the services to an average customer and a more faithful study. As an application to the model described in this paper, the direct implication of service quality as a necessity is instantiated. This allows for the specification of our original goal: to reduce waiting time and balking rate, but also maximize this customer interaction through employee reliability with service, in order to ensure that customers revisit the establishment in the future.

This study aims to model the service at Kung Fu Tea using queuing theory. The store is set up so there will always be one employee at the register and a variable number of employees making the drinks; the number of servers working is dependent on how dense the traffic is for the shift and day. On the weekends, the store typically receives over 200 customers and approximately 350 drink orders.

II. BACKGROUND

A. A. *Modeling Theory*

This paper seeks to implement a model using Markov Chain Queuing as well as a Continuous Time Markov Chain through a birth-and-death system. Continuous Time Markov Chains (CTMC) are stochastic processes that have stationary or homogeneous transition probabilities dependent on only the current state, not past states, which is also referred to as the Markovian Property. CTMCs differ from Discrete Time Markov Chains as they have time distributions rather than set discrete intervals of time.[3] Birth and death processes are CTMCs with states that only transition to the next state or previous state with state transition rates and probabilities found in Figure 1.

$$
\begin{aligned}
v_0 &= \lambda_0, \\
v_i &= \lambda_i + \mu_i, \qquad i > 0 \\
P_{01} &= 1, \\
P_{i,i+1} &= \frac{\lambda_i}{\lambda_i + \mu_i}, \qquad i > 0 \\
P_{i,i-1} &= \frac{\mu_i}{\lambda_i + \mu_i}, \qquad i > 0
\end{aligned}
$$

Figure 1. State transition and probabilities for birth-and-death processes

Queueing theory is an application of a birth and death process defined through an arrival process, service mechanism, and a queue discipline.[1] The arrival process, or inter-arrival process, describes the time between the arrivals of customers in a system and the service mechanism defines the time it takes for the customer to finish a service, where the times can be determined through either a Markovian (exponential), deterministic, or general distribution. Queuing theories typically follow a notation such as Kendall's Full Notation for Queueing Systems of A/S/c/K/N/D where A represents the distribution of time between each arrival to the queue, S the distribution of service times, c the number of servers in the system, K the capacity of the system, N the calling population, and D the queue discipline.[4] Queue disciplines can range from FIFO/FCFS (first in first out), LIFO/LCFS (last in first out), SIRO (served in random order), and GD (general queue

[1] R.B. Cooper "Queueing Theory" (Digital article work style) ACM '81 National Conference, 1981, pp. 120-122.

[2] Bichler, B. F., Pikkemaat, B., & Peters, M. (2020). Exploring the role of service quality, atmosphere and food for revisits in restaurants by using a e-mystery guest approach. *Journal of Hospitality and Tourism Insights*.

[3] S.M. Ross "6.2 Continuous-Time Markov Chains." Introduction to Probability Models (Book style), Tenth ed., Elsevier Inc.: Amsterdam, 2010, pp. 389–390.

[4] S.M. Ross "8.2 Preliminaries." Introduction to Probability Models (Book style), Tenth ed., Elsevier Inc.: Amsterdam, 2010, pp. 498-503.

Optimization of Queueing Model for Local Bubble Tea Shop

discipline). [5] Queues described to be a M/M/1/∞/∞/FIFO would be a Markovian distributed arrival and service rate with a single server, infinite capacity and calling population queue that follows a first in and first out discipline as shown in Figure 2.

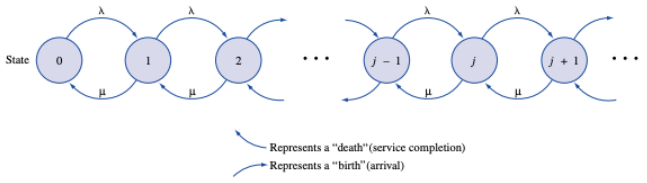

Figure 2. Rate diagram for M/M/1/FCFS/∞/∞ queueing system

Little's Law is a formula for G/G/s queueing systems that outlines parts of a queue system and describes formulas for each. L represents the average number of customers in the system, L_Q the average number of customers in the queue, W the average amount of time customers spend in the system, W_Q the average amount of time customers spend in the queue, and λ_a the average arrival rate of entering customers. The formulas for each part are in Figure 3 below.

$$\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t} \qquad L = \sum_{n=0}^{\infty} nP_n$$

$$L = \lambda_a W \qquad \lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n$$

$$L_Q = \lambda_a W_Q \qquad W = \frac{L}{\lambda_a}$$

Figure 3. Little's Law Queueing Formulas

In order to derive queueing formulas, the birth and death equations or the balance equations must be interpreted for each state in the model in which the rate that the process leaves equals the rate that it enters.[1] For an M/M/k queue, there are a k number of servers and the balance equations can be simplified down to the equations found in the figure below with the long-run probabilities defined.

$$P_0 = \frac{1}{1 + \sum_{n=1}^{k} (\lambda/\mu)^n/n! + \sum_{n=k+1}^{\infty} (\lambda/k\mu)^n k^k/k!},$$

$$P_n = P_0(\lambda/\mu)^n/n!, \quad \text{if} \quad n \leqslant k$$

$$P_n = P_0(\lambda/k\mu)^n k^k/k!, \quad \text{if} \quad n > k$$

Figure 4. M/M/k system stationary probability formulas

Sequential service (or tandem queue) systems characterize queues where customers, on arrival, get serviced by server 1 to completion, then server 2 to completion, and continues to get serviced until the last server. In Figure 5, the tandem queue is portrayed through two servers in which customers enter if server 1 is free, enters service with server 2 after service 1, then leaves the system. Balking can occur in some queueing systems when customers finish service with one server, but cannot move to the next server and leave the system. Balking is not depicted in this figure as customers don't leave if either servers are full. The throughput rate measures the rate at which customers get serviced to completion from the first server to the very last server.

Figure 5. Tandem queue

B.

C. B. Data Collection

To accurately represent the current model of the customer service system at Kung Fu Tea, data was collected for the week of April 18, 2022 to April 24, 2022. For each day, the time, number of transactions, number of items sold, average amount made per check, and total amount earned were collected for each operating hour of the store, which is from 12:00 pm to 8:00 p.m. The study focuses on optimizing service during the store's busiest periods, therefore only the data from Friday, Saturday, and Sunday were used. The number of items signifies the number of drinks sold in the hour, while the number of transactions is the amount of customers that put in an order. The three sets of data from Friday through Sunday were averaged into one table, as shown in Figure 6 below.

Weekend	Time	# Transactions	# Items	Avg. Sales/Check	Sales
Shift 1	12:00-12:59	27.00	44.33	11.02	298.23
	1:00-1:59	23.67	41.00	11.23	268.00
Shift 2	2:00-2:59	27.00	43.00	10.58	286.20
	3:00-3:59	30.67	48.67	10.03	310.10
	4:00-4:59	32.00	48.33	9.56	307.92
	5:00-5:59	32.67	51.67	10.53	343.57
Shift 3	6:00-6:59	24.67	41.00	11.18	275.30
	7:00-7:59	18.67	21.33	7.45	140.08

Figure 6. Data collection for 3 shifts at Kung Fu Tea from April 23rd, 2022

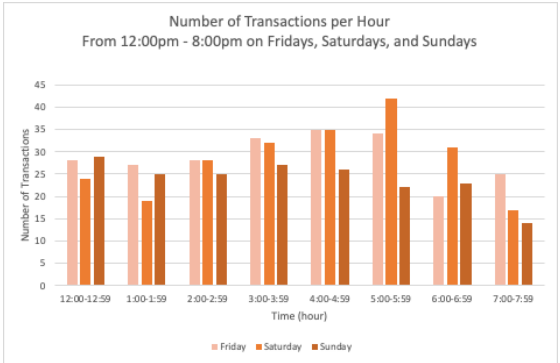

Figure 7. Customers per hour for Friday, April 22nd, Saturday, April 23rd and Sunday April 24th at Kung Fu Tea

[5] W.L. Winston "The Kendall-Lee Notation for Queueing Systems" Operations Research: Applications and Algorithms (Book style), Fourth ed., Hoyt Publishing Services: Belmont, CA, 2004, pp. 1061

Optimization of Queueing Model for Local Bubble Tea Shop

The shifts at Kung Fu Tea are typically divided into three shifts, the first being from 12:00pm - 2:00pm, the second from 2:00pm - 6:00pm, and the third from 6:00 - 8:00pm. Based on Figure 7, it is clear that Shift 2 is the busiest period for the store as there is an increase in the number of transactions, or number of customers, during the time period of Shift 2 and a decrease towards the end of Shift 2 for all three days of the weekend. In order to isolate and optimize the busiest period of time, only the data from Shift 2 was considered as it is the period where the most customers are in the store and the most servers are active.

III. METHODOLOGY

To develop our model, we used a combination of the modeling theory approaches described above in the *Modeling Theory* section. Our model is comprised of a Markov Chain within a Birth-and-Death Process, and the formulations of each will be described further in the sections below:

D. A. Determination of Queue: Birth-and-Death Process

An assumption made in the model is that upon there being 10 customers in the queue, the 11th customer will decide not to join the queue. This refers to the balking, and the balking rate, in turn, is determined by the percentage of customers who come into the establishment and decide not to join the queue. As one of our primary goals with this study is to minimize the balking rate, this means that the average wait time spent in the queue should be reduced. The average wait time is to be determined through type of modeling theory, but the overall queue process could be modeled as below:

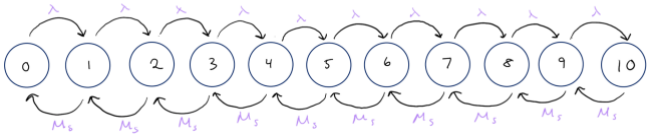

Figure 8. Model of Queue: Birth-and-Death Process

In the figure above, each state within the Birth-and-Death process describes a spot within the queue. Since the 11th customer balks, there are only 10 states included in the queue, plus the state of having 0 customers. Meanwhile, the rate from one spot in the queue to the next (ie. State 0 to 1, State to 2, etc.) is determined by the arrival rate of the customers. The arrival rate was found to be constant on average. Determinations of this constant arrival rate term as well as further considerations will be explored in detail in the next section. For now it can be noted that this arrival rate is 0.7986 customers per minute for the Shift 2 scenario. Meanwhile, the death rate, also known as the rate of customers leaving the queue, is not a constant term and is determined by the average time in the system per customer going through the entire process from waiting in the queue to receiving their drink. This process is modeled through a Markov Chain and will be described in further detail in coming sections. But for now, it should be noted that the Birth-and-Death process being used to model the queue is the baseline of our model, with the rates in between being determined through further modeling theories.

E.

F. B. Arrival and Service Rate Calculation

Using the data from Figure 9, calculations were made so that accurate arrival and service rates could be derived.

	Arrival Rates	Shift 1 (2 servers)	Shift 2 (4 servers)	Shift 3 (3 servers)
	Customers/hr	25.33	30.58	21.67
	Drinks/hr	42.67	47.92	31.17
	Drinks/Cust	1.684210526	1.566757493	1.438461538
	Cust/min	0.4222222222	0.5097222222	0.3611111111
	Drinks/min	0.7111111111	0.7986111111	0.5194444444
	Service Rates			
At register	Drinks/min	1.11	1.11	1.11
At drink making	Drinks/min	0.67	0.67	0.67

Figure 9. Arrival and Service Times

For the numbers used to find the arrival rate, because the model is focused only on Shift 2, all calculations were done using the time period of 2pm to 6pm. The number of customers/hr was calculated by taking the average of the number of transactions between 2-6 pm. The number of drinks/hr was calculated by taking the average number of items/hr between 2-6 pm. The customers/minute and drinks/minute were both calculated by dividing the above values by 60. The number of drinks each customer typically orders was found by dividing the average number of drinks/min by the number of customers/min.

Through repeated extensive observation of the establishment, the service rate for customers ordering at the register averages 0.9 minutes per customer, which means 1.11 customers are processed at the register per minute. The service rate for drink making is 1.4 minutes per drink, or 0.714 drinks made per minute.

G. C. Considerations for Multiple Drinks per Customer and Arrival Rate

One key factor to consider in the model is that many customers do not order just 1 drink. In fact, the number of items a customer orders is often a random occurrence. Other than using probabilities to approximate the number of items a customer orders, this cannot be mathematically calculated. Thus we decided to simplify our model and account for the fact that customers order multiple drinks through the arrival rate of customers itself. It can be recalled that this number is constant and is the birth rate of the queue. Using the data collected, the average number of drinks that a customer orders per shift was determined. Thus when examining Shift 2, the average number of drinks ordered per customer comes out to be 1.5668, which was calcul+ated by dividing the average number of items by the average number of transactions. This number was then multiplied by the average number of customers coming in per minute. The number of customers arriving was given to us with data on the number of transactions for the shift. Therefore for Shift 2 the calculation for arrival rate is as follows:

Optimization of Queueing Model for Local Bubble Tea Shop

λ(arrival rate) = Average # of drinks per customers * Number of customers arriving per minute

λ(arrival rate for Shift 2) = 1.5668*0.5097=0.7986

In summary, we are modeling a situation where each customer only orders one drink, but because the rate of customer arrival was multiplied by the average number of drinks a customer typically orders, the increased rate of customers arriving in the system would account for the fact that a single customer could order more than one drink.

H.

I. D. Markov Chain Process

As mentioned above, the death rate of the actual birth-and-death model is a customer's average time in the drink-making system. To model the service of the customer ordering and getting their drink, we used a Markov Chain as can be seen below:

Figure 10. Markov Chain Process for Drink Service

Here, the queue is the first state within the Markov Chain. Next customers that reach the front of the queue would arrive at the register. The rate in between the queue and the register accounts for the time it takes to order. See the *Arrival and Service Rate Calculations* section for more details. After ordering, the customers proceed to the drink making section. This is the variable part of the system, as the number of servers change depending on what shift it is. Since we are examining Shift 2 specifically, there will be a total of 4 servers: one at the register and three participating in the actual drink making. This is something to keep in mind when the Markov Chain is simulated. Following the drink making, the customer will pick it up and depart the system. In relation to the birth-and-death modeling of the queue, it is important to keep in mind that the Markov Chain Process provides the death rate through the average time spent in the system. Thus it can be reiterated that our system is a Markov Chain Process within a Birth-and-Death Process. The system in its entirety can be shown in Figure 11.

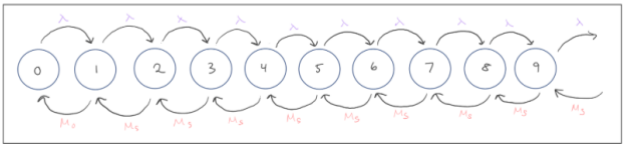

Figure 11. Entire Model of System

J.

K. E. Simulations of Markov Chain

The next step would be to determine the service rate of Shift 2 for the different numbers of customers in the system modeled in Figure 10. This is because the death rate is not constant, and the time a customer takes to get their drink when there are n customers is different from n+1 customers in the queueing system. The service rates were determined through simulating the Markov Chain with Simio, a simulation software. The model was built using a source object for customer arrival, labeled “Entry”, which has an exponential interarrival time of 1.25 minutes. Additionally, a sink object labeled “Exit” was used to represent the departure of the customer from the system. Next, server nodes were implemented, one of them labeled “Register” to represent the drink ordering station for the customer. This has an exponential processing time of 0.9 minutes. There were 3 server nodes used to represent the workers making the drinks, and they had exponential processing times of 1.4 minutes each. We ran these for different amounts of customers, which gave us different amounts of times it takes for them to pass through the system, which were then used to determine the different service rates for the system.

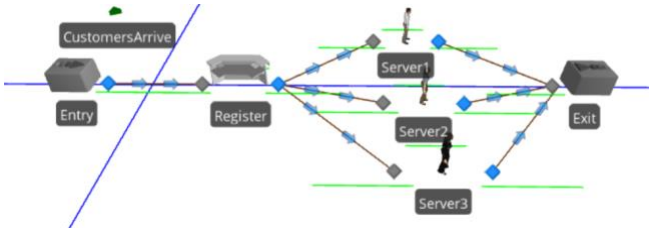

Figure 12. Simio Model for the Markov Chain for Shift 2

Additionally, a Simio model for 2 registers and 2 servers was created (Figure 13), in hopes that a new register could help replace the need and costs for an additional server. The registers each had exponential processing times of 0.9 minutes, and the servers each had exponential processing times of 1.4 minutes each, similar to the section above. This model was also tested with different numbers of customers to determine if the service rates for each are less than the model above in Figure 12.

Optimization of Queueing Model for Local Bubble Tea Shop

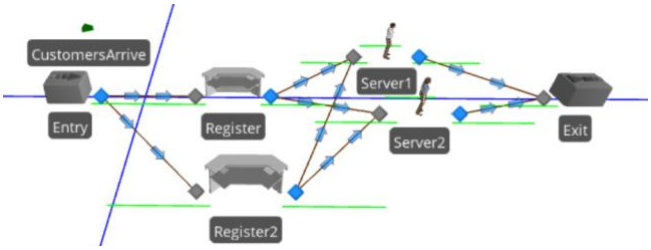

Figure 13. Simio Model for the Markov Chain with 2 Drink-Makers and 2 Registers

L.

M. F. Calculating Probabilities of Birth-and-Death Process

Once the death rates are determined, the overall limiting probabilities for each state can be found. As a reminder, the states for the Birth-and-Death Process represent the number of customers in the queue. Once both the service and arrival rates have been determined, the limiting probability equations were found using the formula:

Rate In = Rate Out

Limiting Probability Equations of Current Model:

$$\lambda P_0=0.492P_1$$
$$\lambda P_1+0.492P_1=\lambda P_0+0.545P_2$$
$$\lambda P_2+0.545P_2=\lambda P_1+0.526P_3$$
$$\lambda P_3+0.526P_3=\lambda P_2+0.499P_4$$
$$\lambda P_4+0.499P_4=\lambda P_3+0.357P_5$$
$$\lambda P_5+0.357P_5=\lambda P_4+0.3415P_6$$
$$\lambda P_6+.2415P_6=\lambda P_5+.2323P_7$$
$$\lambda P_7+.2323P_7=\lambda P_6+.2055P_8$$
$$\lambda P_8+.2055P_8=\lambda P_7+.1697P_9$$
$$\lambda P_9+.1697P_9=\lambda P_8+.1916P_{10}$$
$$0.1916P_{10}=\lambda P_9$$
$$1=P_0+P_1+P_2+P_3+...+P_9+P_{10}$$

1 REGISTER 3 SERVERS

P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	SUM
9.6262E-05	0.00015625	0.000229	0.000348	0.000556	0.001244	0.002910207	0.010005	0.03888	0.182965	0.76261	1

lambda= 0.7986

			Service Rate
7.68748E-05	=	7.69E-05	0.492
0.000201656	=	0.000202	0.545
0.000307625	=	0.000308	0.526
0.000460448	=	0.00046	0.499
0.00072188	=	0.000722	0.357
0.001438112	=	0.001438	0.3415
0.003317927	=	0.003318	0.2323
0.010313842	=	0.010314	0.2055
0.039038972	=	0.039039	0.1697
0.177165355	=	0.177165	0.1916
0.146116134	=	0.146116	

AVG # customers:	9.681426517
Average lambda:	0.7986
Average time in queue:	12.12299839

Figure 14. Probabilities for Current Model (1 Employee at Register, 3 Drink-Makers)

Limiting Probability Equations of Improved Model:

$$\lambda P_0=.664P_1$$
$$\lambda P_1+0.664P_1=\lambda P_0+0.5995P_2$$
$$\lambda P_2+0.5995P_2=\lambda P_1+0.2832P_3$$
$$\lambda P_3+0.2823P_3=\lambda P_2+0.4748P_4$$
$$\lambda P_4+.4748P_4=\lambda P_3+.5310P_5$$
$$\lambda P_5+.5310P_5=\lambda P_4+..3876P_6$$
$$\lambda P_6+.3876P_6=\lambda P_5+.4318P_7$$
$$\lambda P_7+.4318P_7=\lambda P_6+.2796P_8$$
$$\lambda P_8+.2796P_8=\lambda P_7+.3121P_9$$
$$\lambda P_9+.3121P_9=\lambda P_8+.3126P_{10}$$
$$.3126P_{10}=\lambda P_9$$
$$1=P_0+P_1+P_2+P_3+...+P_9+P_{10}$$

2 REGISTERS 2 SERVERS

P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	SUM
0.000738269	0.000888	0.001183	0.003346	0.005628	0.008464	0.017439	0.032254	0.092124	0.235726	0.60221	1

lambda= 0.7986

0.00059	=	0.00059
0.001299	=	0.001299
0.001654	=	0.001654
0.003617	=	0.003617
0.007167	=	0.007167
0.011254	=	0.011254
0.020687	=	0.020687
0.039685	=	0.039685
0.099328	=	0.099328
0.261821	=	0.261821
0.188251	=	0.188251

AVG # customers:	9.289159
Average lambda:	0.7986
Average time in queue:	11.6318

Figure 15. Probabilities for Improved Model (2 Register, 2 Servers)

The linear equations generated using the formula rate in=rate out were inputted into an Excel sheet, as seen in Figure 14 and 15, where Excel Solver was used to calculate P_0 through P_{10}.

The newfound probabilities that were determined for each state were then plugged into Little's law, to solve for waiting time in the queue:

$$L = \lambda_a W_Q$$

We can simplify this formula to:

$W_Q = \frac{L}{\lambda(1-P_0)}$ to account for there not being any customers at the P_0 state. The L term (or the number of customers) was found by multiplying each probability by the number of customers associated with each state.The waiting time in the queue was found for both the old and improved model to be compared. These calculations can be seen below:

Waiting time for current model: 9.6814/0.7986 = 12.123 min

Waiting time for new model: 9.2892/0.7986 = 11.632 min

The balking rate for both the old and new models were also determined by multiplying the average arrival rate with P_{10}, or the probability of being in the state with 10 customers already and having to leave.

Old balking rate: (0.7986)*(0.76261) = 0.609

New balking rate: (0.7986)*(0.60221) = 0.480

IV. DISCUSSION & RECOMMENDATION

We found that the average waiting time for the improved model with 2 registers and 2 drink makers is 11.63 minutes

while the average waiting time for the old model with 1 register and 3 drink makers was 12.12 minutes. Here we see an improvement of 0.49 minutes with our improved model. Additionally, the balking rate decreased from 0.609 to 0.480, which means the rate of customers entering the store and leaving due to high traffic in the store is reduced by 0.129. While these improvements may not be drastic, they do show that even with the same number of servers the current model can be optimized. For real world implementation, the cost of another register being added to the system would have to be factored in as well. However, for a boba shop as popular as Kung Fu Tea it may be worth converting to the improved system to increase customer satisfaction and retain customers. For future research, it would be interesting to see how the system changes when factoring in the fact that different drink makers have different drink making rates. It may also be worth looking into a model that incorporates the probability that a customer will order 1 drink vs 2 drinks vs 3 drinks, etc.

V. Conclusion

In this study, we hoped to optimize and reduce waiting time and balking rate for customers while preserving service quality. The weekend shift from 2-6 pm, or Shift 2, was specifically examined throughout this study since it was the busiest on the three-day weekend period. Our model consisted of a Markov Chain within a Birth-and-Death process. The death rate was obtained using the average time in the system of the Markov Chain for different numbers of customers in the system (1 customer to 10 customers) and implemented to find the limiting probabilities. With these, the time spent in the queue as well as the balking rates could be finalized and compared, and the improved model showed lower values for both quantities. In conclusion, we were able to effectively optimize our model by recommending that Kung Fu Tea have 2 employees at the register and 2 drink-makers.

Appendix

Appendices are attached separately as one file, but listed below.

II. Limiting Probabilities Calculations
III. April 22nd - 24th, 2022 Shifts Data Collection Spreadsheet
IV. Improved Modeling System Data with 2 Register Employees and 2 Drink Making Employees
  A. Infinite amount of customers in the system
  B. 1-10 customers in the system
V. Current Modeling System Data with 1 Register Employee and 3 Drink Making Employees
  A. 1-10 customers in the system
VI. Improved Model with 2 Register Employees and 2 Drink Making Employees
VII. Current Model with 1 Register Employee and 3 Drink Making Employees

Acknowledgment

We wish to acknowledge the help provided by the University of Virginia Stochastic Decision Model Teaching Assistant team: Christos Chen, Aatmika Deshpande, Caton Gayle, and Aparna Ramanan. We would also like to show deep appreciation to our supervisors Professor Robert Riggs and PhD Student Instructor Aram Bahrini who helped us finalize our project.

References

[1] Bichler, B. F., Pikkemaat, B., & Peters, M. (2020). Exploring the role of service quality, atmosphere and food for revisits in restaurants by using a e-mystery guest approach. *Journal of Hospitality and Tourism Insights*.

[2] R.B. Cooper "Queueing Theory" (Digital article work style) ACM '81 National Conference, 1981, pp. 120-122.

[3] S.M. Ross *Introduction to Probability Models* (Book style), Tenth ed., Elsevier Inc.: Amsterdam, 2010.

[4] W.L. Winston *Operations Research: Applications and Algorithms* (Book style), Fourth ed., Hoyt Publishing Services: Belmont, CA, 2004.