Singapore Healthcare Management 2024

Optimizing Rostering for NUHS Group Contact Centre to Maintain Customer Satisfaction using A Data-driven Simulation Approach

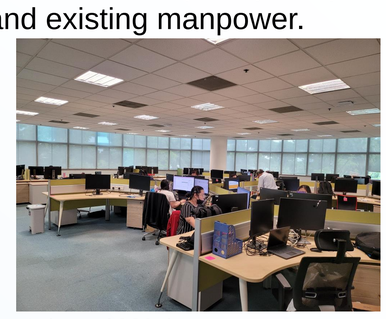
Xiao JIN, Institute of Operations Research and Analytics
Chung-Piaw TEO, Institute of Operations Research and Analytics
Jessyln NG, National University Health System
Jasmine HONG, National University Health System
Yi Zhang TAN, National University Health System

Introduction

- Due to the nature of calls for healthcare industries, service availability is crucial and one of the key performance indicators tracked is abandoned rate of hotlines.
- As calls remained a critical touch point that is heavily dependent on human agents, there was a need to validate that NUHS Group Contact Centre (GCC) is operating optimally with the implementation of digital solution and existing manpower.

Hotline / Channel	Abandoned Rate	Waiting in Queue	Longest Waiting Time
Appointment	3 %	16	00:03:33
Dental	0 %	0	00:00:00
GP	0 %	0	00:00:00
General Enquiries	1 %	1	00:02:18
Live Chat	0 %	-	00:00:45

A snapshot of the live wallboard in GCC. Red square highlighted a surge of arrival, causing abandoned rate to rise.



A picture of GCC site

- GCC operates 24/7. However, our focus for this study is from 8:00am to 5:30pm (19 half-hour slots), Mondays to Fridays. We aim to answer the following question:

To keep abandoned rate of Appointment (AP) and General Enquiry (GE) lines below 5%, what is the minimal number of staff required for each half-hour slot?

Challenges:

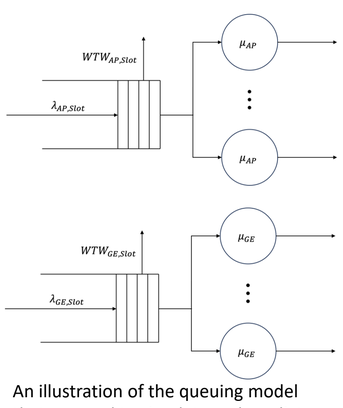
1. **Variable Caller Patience:** Based on data and statistical inference, it has been observed that callers' Willingness-To-Wait (WTW) fluctuates over time.
2. **Stochastic Arrival Patterns:** Call arrival times are stochastic and unpredictable, with varying likelihoods of sudden surges throughout the day.
3. **Interference from Past Decisions (Snow-ball effect):** The performance of staffing decisions for a half-hour depends on the queue status at the beginning of that slot and is affected by all previous staffing decisions, arrival patterns, and caller patience levels, making it difficult to obtain a representative sample.

Methodology

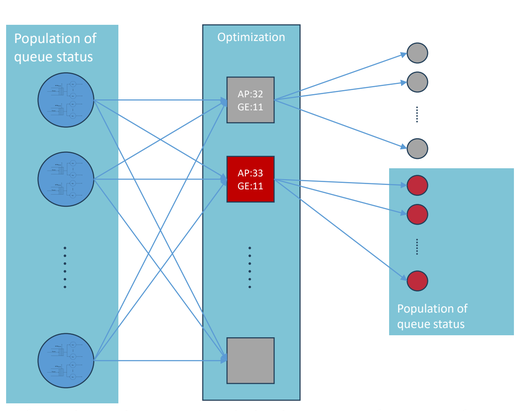
- To model customer abandoned behavior and infer latent Willingness-To-Wait (WTW), we rely on the relationships among three random variables.
  - X denotes the WTW of a customer,
  - Y denotes the waiting time required to be served if the customer does not abandon the queue.
  - Z denotes the actual queuing time.
  - $Z = \min\{X, Y\}$. Z and Y are observed from CSQ data. By plotting $\frac{P\{Y>x\}}{P\{Z>x\}}$, we can infer the distribution of WTW:

$$P\{X > x\} = \frac{P\{Y > x\}}{P\{Z > x\}}$$

- To estimate abandoned rate of each half-hour slot, a simulation model is built based on $M_t/E/n \; - \; G_t$ queue system:
  - M_t: Interarrival follows exponential distribution which changes every 30 mins. Arrival rate is learned from the call queue (CSQ) data
  - E: Service time follows Erlang distribution due to multiple phases during a call. The service time distribution is assumed unchanging. The parameter is learned from CSQ
  - n: Number of staff allocated to a line
  - G_t: WTW follows general distribution learned from CSQ data
- To derive the optimal staffing level for each slot, we apply a sequential simulation-based optimization method that fully captures the complexity of the queue status.
- To address forecast risks, the performance of each staffing plan is estimated based on a quantile-level analysis.



An illustration of the queuing model that we used to simulate and predict abandoned rate.

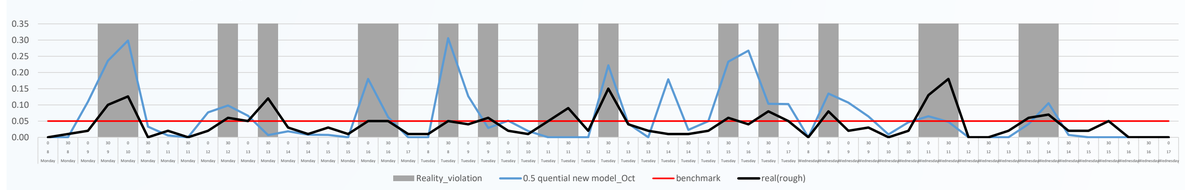


An illustration of the simulation-based optimization process for one half-hour slot. We inherit n system statuses from previous slot simulation then simulate the system for 30 minutes for each of the n statuses, conditioned on each possible staff allocation. The optimal staffing allocation is selected based on the quantile performance among the n simulation trials. These n trials, conditioned on the optimal staff allocation, will then be inherited as the starting population for the next slot simulation.

Result

Validation:

- In sample performance:



Using CSQ data and actual manpower staffing in Oct 2023 (Mon to Wed), we simulated the abandoned rate (for both AP and GE lines at 50% quantile level) and compared the results against actual abandoned rate recorded. Grey area highlights the 18 slots where abandoned rate >5%. Among 18 slots, the simulation (black plot) highlighted that abandoned rate is >5% in 12 slots successfully, hitting a 66.7% correct alarm rate.

- Field experiment: the optimal staffing can maintain abandoned rate <5%

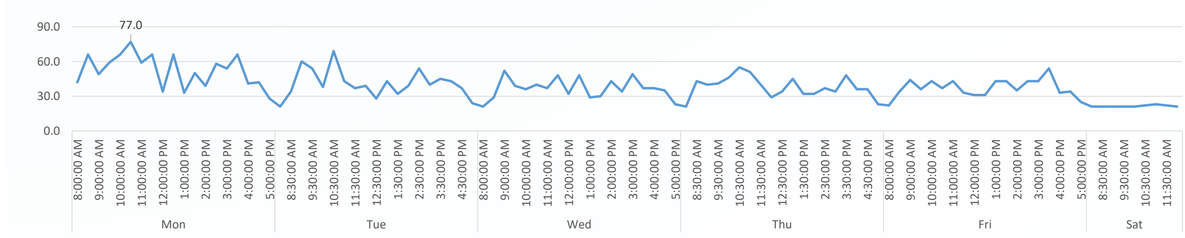
	Abandoned Rate for AP Line	
	Week 1 (1 to 5 Jan)	Week 2 (8 to 12 Jan)
Mon	-	2.8%
Tue	2.7%	3.2%
Wed	3.2%	1.8%
Thu	3.4%	3.5%
Fri	1.9%	2.5%
Overall	**2.8%**	**2.8%**

	Abandoned Rate for GE Line	
	Week 1 (1 to 5 Jan)	Week 2 (8 to 12 Jan)
Mon	-	5.1%
Tue	3.4%	3.6%
Wed	4.4%	2.9%
Thu	4.8%	4.0%
Fri	4.6%	4.2%
Overall	**4.2%**	**4.0%**

CSQ data for Dec 2023 was used to train simulation, predict the caller's arrival and derive the optimal staffing on 95% quantile level. The suggested staffing was then used by GCC for rostering in the first two weeks of Jan 2024. Based on the results above, we can conclude that the model is effective as abandoned rate is maintained <5% in all but one slot.

Applications of Simulation Model:

1. This function is implemented for AP and GE lines to find out minimum staffing level required to maintain abandoned rate at <5%. Through the study, GCC found following insights for an efficient rostering:
   - ✓ Adjusted staff's lunchtime to cater to peak hours
   - ✓ Limit staff leave and trainings on Mondays, post public holidays and long weekends
   - ✓ Use IVR to encouraged callers to call after peak hours



The graph shows the total staffing required under the optimal staff scheduling for each half-hour slot for Jan 2024 at 95% quantile level. The plot suggested that the peak hour is likely to happen during 10:30am~11:00am, Monday and the minimum staff required to ensure abandoned rate is <5% is 77. It is noteworthy that the number of staff required would decrease significantly after Monday.

2. Evaluate existing rostering plan.

Plot of an existing staff scheduling and alarm suggest by simulation. The red slots highlight that abandoned rate would not be maintained at <5% for 95% of the time. Though this, GCC is able to review the roster and add more staff as suggested.

Conclusion

1. Our data-driven simulation method provides an end-to-end solution from call data to optimal rostering. It's easy to implement and can be extended to other contact centre cases such as NUCOHS and NCIS lines managed by GCC.
2. Given arbitrary rostering, the trained model can predict worst case abandoned rate of each slot with accuracy and can alarm staffing deficiency of existing rostering plan.
3. The optimized rostering plan proved to be reliable and effective in keeping abandoned rate <5%.

NUHS National University Health System

NUS National University of Singapore