### **RESOURCE SCHEUDLING IN NON-STATIONARY SERVICE SYSTEMS**

Samira Shirzaei Jeffrey S. Smith

Industrial & Systems Engineering Department Auburn University 3301 Shelby Center Auburn, AL 36849, USA

### ABSTRACT

We focus on a service system in which the customer arrivals are non-stationary and our goal is to determine a server staffing schedule that ensures that arriving customers do not experience long and/or unpredictable queue times. An airport ticket counter is an example of such a system. Passengers arrivals are nonstationary, yet arriving passengers do not wish to wait in long lines to check into their flights. Moreover, unpredictability is a significant issue in these environments as it often forces passengers to arrive earlier than necessary "just in case." Unfortunately, we rarely know the precise form of the arrival process and must use observed samples to set the staffing policy. We show through a case study that simulation combined with a specialized input analysis tool can be used to determine good staffing policies in these environments.

## **1** INTRODUCTION

Our goal is to optimize a service system operation like a check-in counter in an airport, by focusing on the staffing levels to best control the customers' waiting times. We assume that passenger arrivals are nonstationary, but we do not know the precise form or parameters of the arrival process. It is clear that if the service rate is significantly larger than the maximum arrival rate, customers will generally experience a small amount of waiting time, but such overcapacity is expensive in terms of resource costs. While the expected queue time is important, we are more interested in the predictability/variance of the queue time as described by Smith and Nelson (2015) and our objective is to determine a resource schedule that can make the waiting times appear *practically stationary* when the arrivals are non-stationary.

Input analysis is arguably one of the most expensive steps in most simulation studies and is essential to a successful simulation (Law 2009). An important step in input modeling is the assessment of data being independently and identically distributed (IID). While this is straightforward when modeling stationary stochastic processes, it becomes more challenging when the stochastic process follows a non-stationary pattern where the probability distribution or its parameters depend on time (Ansari et al. 2014). They proposed the Histogram and Rates for Input Analysis (HistoRIA) as a tool to facilitate input modeling. Smith and Nelson (2015) used a time bucket method for estimating virtual waiting time for the customers arriving to the queue of check-in section of the airport with non-stationary input arrivals. They showed that averages of waiting times within time buckets are more relevant to individual customers than the overall average. One factor that strongly affect this waiting time is number of assigned servers in each time bucket. If we can determine them properly we will be able to control customer's satisfaction dramatically. It is obvious when we have infinite number of servers the waiting time is in the least amount but the cost of having infinite servers is too much and we are interested to reduce the cost of using them.

We focus on scheduling server levels in our case study with the goal of cost minimization under customer satisfaction constraints. In related work, Feldman et al. (2008) developed a method to determine

appropriate staffing levels in call centers with the goal to achieve targeted time-stable performance and they assumed sinusoidal arrival-rate function  $\lambda(t)$ . Jennings et al. (1996) considered a multi-server system with general nonstationary arrival and service time process. They developed an approximate procedure based on a time-dependent normal distribution where the mean and variance are determined by infinite-server approximation. Green et al. (2007) reviewed queueing-theory methods for setting staffing requirements in service systems with time-varying customer's demand. They showed how to adapt stationary queueing models for use in nonstationary environments. Depending on level of targeted quality of service and service times, they discussed which method should be used and how must be modified. Whitt (2007) discussed methods to encounter with time-varying demand to set staffing levels in call centers, he showed when and why, and what to do when each of those methods fail. Izady et al. (2011) tried to set minimal medical staffing levels for reducing patient's waiting time in the presence of complexities like time-varying demand, multiple types of patients, and resource sharing. Their proposed staffing algorithm relies on infinite server networks to compute the resources' time dependent workload and highlights their ability in modeling complexities like multiple types of customers. These articles assumed arrival rate follows pre-determined distributions, sinusoidal arrival-rate function, but here we have only a set of sample data to work with and do not know the process characteristics. Furthermore, they used some approximations to obtain the appropriate resource schedule such as normal approximation and delay probability approximation and those estimations were affected by the size of parameters but in this paper we are interested to have less limitation in the process and system's characteristics and perform staffing levels optimization with experimentation in simulation software with closer assumptions to make the model more realistic and observing practically stationary waiting time.

## 2 INITIAL MODEL

Our basic model's characteristics are similar to the one used in Smith and Nelson (2015), which has a single-server with non-stationary arrival process and a time-dependent agent capacity. The primary difference here is that they used this model to measure passengers' time in system and we are using the model to optimize the facility staffing levels so that arriving passengers can better predict waiting times.

## 2.1 Input Analysis for Arrival Data Set

We start with the passenger arrival data for 5 days from an airport check-in counter. In general, we do not have information about what kind of stationarity we can consider for our dataset. In some cases the data are stationary session-to-session, month-to-month, week-to-week, day-to-day and so forth. We used the HistoRIA tool (Ansari et al. 2014) to characterize the sample arrival data. This is a graphical tool that helps the user assess the stationarity of a process over time. To simplify our initial work, we assume our data are stationary on a day-to-day basis and that we can view the data as including 5 observations of a "generic day."

This tool gets all the passengers' intervals for each day as the input and after calculating the interarrival times, converts them to corresponding timestamps. Using these timestamps the tool gives the HistoRIA plot as the output. For drawing the Historia plot, we should determine time block size which shows we are dividing the day to how many smaller intervals. The lengths of these time buckets are important because they must be short enough that arrivals during the interval see essentially the same system load, but long enough so that it is almost certain that there will be arrivals during each interval (Smith and Nelson 2015). In addition, we should determine the number of studying days which is in this case five and we are considering how many hours in a day. Figure 1 shows two sample HistoRIA plots for the given dataset. In the first HistoRIA plot, each time block represents one hour and in the second plot, each represents one half an hour. Therefore, there are 24 and 48 time blocks, respectively. In each time block, the gray area is the estimated arrival rate ( $\lambda(t)$ ). As it can be seen in the figure, the average arrival rate in hourly blocks is almost equal to average of the corresponding 30-minute blocks' arrival rates, so it seems reasonable to

choose our time bucket equal to one hour for this data set. We used the estimated arrival rate in 1-hour time bucket as the input for our airport check-in model in Simio.

As mentioned above our goal is determining number of servers as the function of time, we consider the airport check-in system with nonstationary passenger arrivals and we focus on the system from both airport manager and customers' views, so with having optimal staffing levels they all will be satisfied. If there were no server constraints there would be no waiting time in the system, but this would almost certainly waste capacity since the arrivals are stochastic and nonstationary. Figure 2 shows a dynamic status plot for one replication of the model with infinite server capacity assumption. This plot shows the number of passengers in the queue (NIQ) and overall time in system (TIS) over simulated time. There are no passengers in the queue, the average time in queue is zero (both TIQ and NIQ remained zero during simulation time) and the average time in system in hourly buckets is almost equal to mean service time which is 3 minutes – exactly what we expect. Unfortunately for flyers, infinite server capacity is not practical.



#### 1- hour time block



#### 30-minutes time block

Figure 1: HistoRIA plot for passengers' interval data set.

Shirzaei and Smith



Figure 2: status plot for model performance metrics with no server constraint capacity.

## 2.2 **Problem Description**

From a customer service perspective, the best system is one that has lots of servers so that no arriving customer waits. However, this is generally not practical from a cost perspective. Our goal is minimizing cost of using servers with minimum customer waiting time and customers have the perspective about amount of waiting time in the queue before arrival to the service system – i.e., the waiting time is *predictable*. Note that the desire for short queue times and predictable queue times are two separate objectives and our focus here is on the predictability. As we encounter non-stationary input data without information on the precise distribution functions, the problem of staffing level scheduling is not straight forward. Customers are interested to know when they should leave their house to make their flight. In that paper they proposed a method to determine passengers' TIS for virtual waiting time estimation. Here we are going to look at the problem from both costumer and airport management perspective. Airport management likes to reduce the cost of using resources and increase the customer service level and customers like to have less waiting time in the check-in queue and would be able to have prediction about this time. We look at this problem as optimization model. This model is given by

$$\min\left[c\int_{t=0}^{24} s(t)dt\right],$$
$$mr\left[TIO(t) < a\right] > b.$$

Subject to:

Where *c* is the unit resource cost, s(t) is the number of resource units assigned at time *t* (i.e., the schedule), TIQ(t) is time-in-Queue at time *t*, and *a* and *b* are constants related to waiting time and service level, respectively. For example, if at an airport check-in counter the probability of customers' TIQ being less than 15 minutes is more than 90 percent, we assume that the customers will be satisfied. The objective function is minimizing the cost of using resources, the decision variables is determined by s(t) which are the staffing levels in time *t* that we are going to determine, and the constraint is service level which we control that with s(t). While the arrival data are nonstationary during the day we use our model to determine the agent resource schedule that will yield approximately stationary TIQ process – what we call *practically stationary*.

Figure 1 showed number of arrivals is changing dramatically during the day, the range of rates is 716 arrivals with minimum of 10 and maximum rate of 726 customers per hour. The manager's goal is to control the arriving customer waiting using the resource schedule – setting the number of servers as the function of time t. We do not consider customer abandonment because of system type but it can be important in general service systems. However, in most of those systems if we schedule the required staffing levels in the way that customers experience a practically stationary system, the probability of abandonment will decrease meaningfully. Although our model is time continuous, we make staffing level changes at discrete times by dividing the day into smaller intervals of length one hour and for evaluating customers' waiting time we consider the average waiting time in that time bucket.

## **3** EMPIRICAL APPROACH

To clarify the importance of having appropriate staffing levels in nonstationary processes, we show some examples of having predefined resource schedule without considering the non-stationarity of customer arrivals. For the arrivals we obtained from input analysis in previous section, we run the simulation model in Simio. The maximum number of arrivals is in 12th time bucket and it is about 726 customers. We assumed the service rate is 20 customers per hour so if manager decides to have constant staffing levels in whole day say, 35 servers, we should see small amount of waiting time in the queue, in this case maximum average passenger's time in queue in hourly bucket is about 3 minutes and 40 seconds. For obtaining these average waiting time in each bucket, we used same approach in Smith and Nelson (2015), we need to consider time-in-queue ordered by the time of customer arrival. We export the required data from Simio to external program to do more statistical calculation and plot these average time-in-queues in each time bucket such the servers are idle with small utilization. We decreased these constant staffing levels to 25 and 15 servers and observed its effect on average customers' waiting time in each time buckets . The results are shown in Figure 3.

When we decreased the service level to 25, the passenger's average time in queue increased to about 80 minutes in the rush hours of the day, however there are noticeable number of time buckets with no customer in the queue that affects the cost of using resources. With having 15 servers, the system will not be able to serve the customers and at the end of the simulation time there are still passengers in the queue, there is a time bucket with average waiting time more than 400 minutes and the simulation did not show the result for time bucket after 16<sup>th</sup> one, because the system is not stable any more.

From above discussion it seems it is impractical to have constant resource levels during the day in such a non-stationary arrival process. In some parts of the day system is overstaffed while others are understaffed. In overstaffed parts the manager is paying extra for using resources and in understaffed ones customers will experience long waiting time before getting the service. So, we should have appropriate staffing levels to optimize the cost of using resources with regard to having practically stationary waiting time from the passengers' perspective.



Constant 15 servers in each time bucket

Figure 3: Plot for passengers' average Time-in queue (in minute) and Number-in-queue.

## 4 DISSCUSSION ABOUT POSSIBLE SOLUTIONS

We used a manual optimization approach in Simio and considered staffing levels and average passenger waiting time in each time bucket at the same time. As we mentioned, customers would like to have predictable waiting time in the queue regardless of the time of the day. We arbitrarily assume if a passenger's average waiting time in the queue is less than 10 minutes, they will be satisfied. With considering this constraint we are trying to have minimum staffing levels in each time block that will lead to minimum cost of using resources which is our objective function. For calculating the average customers' waiting time in each time block in Simio, we defined a Tally statistics for each of the defined 24 time blocks. We need to tell the model to store each TIQ for each entity in the correct Tally statistic. We need to calculate the difference between passenger's arrival time and time before starting the service so we executed add-on process when an entity enters the before processing status in the agents object. Refer to Figure 4 in their work for details.

In Simio, for tracking these average hourly time in queues, we designed 24 responses and also we added the response to record agents' utilization over the day. We applied these responses and the *work schedules* table to find the optimal staffing levels for our check-in counter model with non-stationary arrivals.

The basic algorithm for setting the staffing level is outlined below.

- 1. Put the observed arrival time data in the HistoRIA tool to determine the time bucket length.
- 2. Use the obtained arrival rates from step 1 in the Simio model in associated time blocks.
- 3. Set staffing levels in time buckets based on an initial guess according to arrival rates.
- 4. Set the targeted maximum waiting time (a) equal to 10 minutes and the probability of satisfying this constraints (b) equal to 0.9 (or other constants).
- 5. Use the Simio model to calculate the time-in-queue for each time bucket.
- 6. Check in which time blocks, resource levels are satisfying the constraints in optimization equation in section 2.2, for those have been satisfied reduce the resource level one unit and go to the step 5. For those have not been satisfied, increase the staffing level one unit and go to step 5. For staffing levels that we decreased them one unit and the response has been worsened, change their values to the previous ones and keep them unchanged.
- 7. If all the staffing levels are in unchanged state, return the solution as "good."

In Figure 4 we show the solution for service levels scheduling resulting from simulation model. For obtaining these server assignments we used the initial guess according to arrival rates we obtained from input analysis and average service time distribution. According to this initial guess, we ran the simulation model and observed if these levels satisfy our constraints and our goal for having minimum number of servers which leads to minimum cost (for this purpose, if we observed the bucket with too small average time-in-queue, we can decrease our initial guess for its server assignment and see what will happen for the average waiting time with this new number of servers). We continued this trial and error procedure to achieve to the best assignment (when we cannot decrease the number of assigned servers in each bucket and the waiting times still satisfy the constraint). For these levels of resources, the servers' utilization over the day is about 97.208 % and the maximum average waiting time was observed in 7<sup>th</sup> time bucket which is about 9.49 minutes that meets our constraint about passenger's satisfaction. The total number of resources we used in 24 buckets of the day is 316. If we use constant 35 servers in each bucket, in total we will use 840 resources in all 24 time buckets and in 13<sup>th</sup> one we observed the average time in queue about 30.8 minutes and the agents utilization is about 43.09%.

As we determined in the constraint of problem's model, we should set up parameter *b* which shows what percentage of our passengers are satisfied, so we cannot decide about staffing levels strictly according to average waiting time in each bucket. For this purpose we considered the 90 percent percentile for each of these responses in the Simio. We ran 300 replications of our model and used SMORE plot as tool to help making the decision. The result shows that the staffing levels are slightly different than the case we considered only average of waiting times in the buckets. Figure 5 shows two of resulting SMORE plots (we arbitrary choose these four buckets) for average time-in-queue responses and updated staffing levels for our check-in model.

Using the updated staffing levels, in total we need 329 server-hours during the day and the servers' utilization is about 93.65 %. In this approach we know in at least 90 percent of the situations, passengers will not experience waiting time more than 10 minutes in all time buckets. So, if we'd like to observe percentage of customers that their waiting time was less than 10 minutes more than 90 percent, we must have more servers in some time buckets in comparison to the situations we'd like to observe average waiting time in each time block less than 10 minutes.

For the goal of experiencing practically stationary time in queue from customers' perspective, we can conclude that when arrival rates are changing dramatically we cannot observe same resulting average time in queues in time buckets because by adding or removing one server we will observe totally different average waiting time and in some buckets the range can change about 6 minutes. Instead we can determine the upper bound for average time in queue. In the time buckets with low arrival rate, the time in queue will be low even by having only one servers, which is good from both manager and customers' perspective, so we will not concentrate at staffing level schedule for those too much. when the passengers are going to plan

when should they leave the house to be able to make their flight, they can have prediction for average amount of this waiting time.

1		L .					
	۹	Start Time	Duration	End Time	Value	Cost Multiplier	Description
		12:00 AM	1 hour	1:00 AM	1	1	
		1:00 AM	1 hour	2:00 AM	1	1	
		2:00 AM	1 hour	3:00 AM	1	1	
		3:00 AM	1 hour	4:00 AM	1	1	
		4:00 AM	1 hour	5:00 AM	4	1	
		5:00 AM	1 hour	6:00 AM	6	1	
		6:00 AM	1 hour	7:00 AM	9	1	
		7:00 AM	1 hour	8:00 AM	20	1	
		8:00 AM	1 hour	9:00 AM	28	1	
		9:00 AM	1 hour	10:00 AM	28	1	
		10:00 AM	1 hour	11:00 AM	35	1	
		11:00 AM	1 hour	12:00 PM	35	1	
		12:00 PM	1 hour	1:00 PM	37	1	
		1:00 PM	1 hour	2:00 PM	24	1	
		2:00 PM	1 hour	3:00 PM	21	1	
	۲	3:00 PM	1 hour	4:00 PM	18	1	
		4:00 PM	1 hour	5:00 PM	14	1	
		5:00 PM	1 hour	6:00 PM	12	1	
		6:00 PM	1 hour	7:00 PM	7	1	
		7:00 PM	1 hour	8:00 PM	5	1	
		8:00 PM	1 hour	9:00 PM	3	1	
		9:00 PM	1 hour	10:00 PM	2	1	
		10:00 PM	1 hour	11:00 PM	2	1	
		11:00 PM	1 hour	12:00 AM	2	1	
	*						

Figure 4: Staffing levels for check-in counter at the airport.

The above discussion gives us good information about the average time-in-queue by hour and we used replications to computes confidence intervals. It would be risky for the manager to decide about staffing levels only based on the average waiting time in queue because it contains variability and customers cannot rely on only average waiting time to plan their leaving time from house for the same reason. We can calculate the standard deviation of these mean time-in-queues using the reported confidence interval half-width from the SMORE plot but this standard deviations give us information only about bucket means waiting time in queue. We used the approach of Smith and Nelson (2015) to calculate individual standard deviations. We ran 300 replications of our model and used an external Python program to aggregate the data from these files for further analysis. Table 1 shows outputs of these analysis. For each time block we have the mean and standard deviation of passengers' time-in-queue and number of observations. As mentioned before we have the highest arrival rate in 12<sup>th</sup> time bucket, and in this table we have highest sum of number of observation in these 300 replications in that bucket which is 218160 observations overall.

The results show that in all time buckets, average waiting time in time blocks calculated with simulation and statistical method in Python are the same but simulation underestimated the standard deviation and the passengers cannot plan only with considering average waiting time's standard deviation. When we consider the mean of time in queues, we conclude that we can determine the upper bound or range for

passengers' waiting times for practically stationary concept. We cannot say the means are exactly the same but the travelers can anticipate to see predetermined maximum waiting time in the queue.

Simula	Simulation-reported Results			Computed Results			
Hr	Mean	Std. Dev.	Obs	Mean	Std. Dev.		
12:00	1.986	2.47	2920	1.986	3.84	-35.67%	
1:00	4.609	4.15	4478	4.609	6.00	-31.08%	
2:00	0.528	0.99	3556	0.528	1.68	-41.07%	
3:00	1.507	1.82	7631	1.507	2.82	-35.46%	
4:00	2.469	2.28	22492	2.469	3.30	-30.90%	
5:00	2.011	2.16	37532	2.011	2.76	-21.73%	
6:00	2.758	2.06	60226	2.758	2.88	-28.47%	
7:00	3.533	2.18	135368	3.533	2.88	-24.31%	
8:00	4.598	2.93	179089	4.598	3.24	-9.56%	
9:00	3.709	3.38	165568	3.709	3.66	-7.65%	
10:00	2.102	2.39	210121	2.102	2.58	-7.94%	
11:00	3.073	2.76	218160	3.073	3.00	-8.69%	
12:00	1.776	2.39	203074	1.776	2.70	-11.48%	
13:00	1.321	2.51	119801	1.321	2.76	-9.05%	
14:00	1.567	2.31	120407	1.567	2.58	-10.46	
15:00	2.858	3.14	97469	2.858	3.42	-8.18%	
16:00	4.565	4.28	89688	4.565	4.62	-7.35%	
17:00	2.939	4.32	67307	2.939	4.80	-10.00%	
18:00	3.729	5.96	37210	3.729	6.30	-5.39%	
19:00	4.066	7.25	30219	4.066	7.80	-7.05%	
20:00	4.728	10.86	11845	4.728	11.22	-3.21%	
21:00	4.604	9.73	8891	4.604	10.20	-4.61%	
22:00	3.045	6.98	7579	3.045	7.56	-7.67%	
23:00	3.358	3.36	7361	3.358	3.60	-6.66%	

Table 1: Comparison of the simulation-reported and computed results for means and standard deviations. The negative deltas show simulation underestimate the standard deviation.

# 5 CONCLUSIONS

In real customer service systems, arrival processes are often non-stationary. This nonstationary can make resource planning difficult as the system has the competing objectives of customer service and resource cost. This issue will be harder when we do not have precise information about input data (which we generally will not). This paper illustrates the situation with a case study where a manager must determine the staffing levels in the check-in counter of the airport with non-stationary passenger arrivals. We developed and demonstrated a simulation-based approach to determine appropriate staffing levels from manager and passengers' view and define the problem as optimization model to achieve predictable waiting time results. With these result passengers would be able to schedule their leaving time from the house to the airport without concerning about being late and airport managers are certain they have satisfied their customers with minimum possible cost of using resources. A logical next step for this work would be to generalize the methodology so that it would be applicable in other similar situations and to refine the time-bucket selection/identification method to potentially automate this process.

Shirzaei and Smith



Figure 5: SMORE plot for two hourly buckets and tabulation of staffing levels.

### REFERENCES

- Ansari, M., Negahban, A., Megahed, F. M., and Smith, J. S. 2014. "HistoRIA: A New Tool for Simulation Input Analysis". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 2702–2713. Piscataway, New Jersey: IEEE.
- Feldman, Z., Mandelbaum, A., Massey, W. A., and Whitt, W. 2008. "Staffing of Time-Varying Queues to Achieve Time-Stable Performance". *Management Science* 54(2):324–338.
- Green, L. V., Kolesar, P. J., and Whitt, W. 2007. "Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System". *Production and Operations Management* 16(1):13–39.
- Izady, N., and Worthington, D. 2012. "Setting Staffing Requirements for Time Dependent Queueing Networks: The Case of Accident and Emergency Departments". *European Journal of Operational Research* 219(3):531–540.
- Jennings, O. B., Mandelbaum, A., Massey, W. A., and Whitt, W. 1996. "Server Staffing to Meet Time-Varying Demand". *Management Science* 42(10):1383–1394.
- Law, A.M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. New York: McGraw–Hill, Inc.
- Smith, J. S., and Nelson, B. L. 2015. "Estimating and Interpreting the Waiting Time for Customers Arriving to a Non-Stationary Queueing System". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 2610–2621. Piscataway, New Jersey: IEEE.
- Whitt, W. 2007. "What You Should Know about Queueing Models to Set Staffing Requirements in Service Systems". *Naval Research Logistics* 54(5):476–484.

## **AUTHOR BIOGRAPHIES**

**SAMIRA SHIRZAEI** is a doctoral student at Department of Industrial and System Engineering at Auburn University. She received her Master's and Bachelor of Science degrees both in Industrial Engineering from Iran University of Science and Technology, Iran. She was instructor in Industrial Engineering Department in University of Sistan and Baluchestan, Iran. Her research interests are in simulation, operations research, and supply chain management. She is vice-president in INFORMS student chapter in Auburn University and her email address is szs0165@auburn.edu.

**JEFFREY S. SMITH** is the Joe W. Forehand Professor of Industrial and Systems Engineering at Auburn University. His research and teaching interests include simulation modeling and analysis, manufacturing system design, and analytics for operations. He has served as the WSC Business Chair (2010) and General Chair (2004) and is currently on the WSC Board of Directors. He has a BIE from Auburn University and a MS and PhD (both in Industrial Engineering) from Penn State University. His email and web addresses are: jsmith@auburn.edu and http://jsmith.co.