INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Minimizing Customer Waiting Costs for Rental Vehicle Providers using Threshold Reservation Policies

Jennifer A. Pazour
Debjit Roy

W.P. No. 2012-12-05
December 2012

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD – 380015
INDIA

Minimizing Customer Waiting Costs for Rental Vehicle Providers using Threshold Reservation Policies

Jennifer A. Pazour
Debjit Roy

Abstract: Vehicle rental providers offer differentiated services to reserve and walk-in customers. In this research, we study one such service differentiation strategy, a vehicle threshold policy, which is to hold vehicles for reserve class customers in anticipation of their future arrivals. To consider the impact that vehicle threshold policies have on reserve and walk-in customer waiting times, we model a rental depot as a multi-class non-work-conserving semi-open queue with stochastic inputs. For exponential and deterministic service time distributions, we identify the optimal threshold quantity for stationary customer arrivals using closed-form expressions for the expected waiting times of both customer classes. For non-stationary customer arrivals, we develop different threshold policies and analyze their performance using a detailed simulation model. Through numerical testing, we provide insights into recommended threshold policies that can be applied to improve the profitability of a rental provider.

Keywords: Transportation; Vehicle rental system; Customer service differentiation; Semi-open non-work conserving queue; Priority threshold queuing systems.

1. Introduction

Vehicle rental providers operate a fleet of vehicles that are rented to customers who are temporary in need of a vehicle for a fee. Vehicle rental markets are growing worldwide, with the United States' market witnessing an 8.8% increase in revenue growth in 2011 from $20.6B in 2010 [2]. Notable large rental providers include Enterprise, Hertz, and Avis with a fleet size of 920,861, 320,000, and 285,000 vehicles, respectively.

Rental providers serve a wide range of customer needs and rental periods (generally ranging from a few hours to a few weeks). Customers rent vehicles for business travel, for leisure travel, for a replacement vehicle due to accidents or vehicle maintenance, and for use as a primary vehicle. Customers, who plan their travel in advance, reserve vehicles whereas customers with last-minute

[1] Jennifer A. Pazour, Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL 32816, Email:jennifer.pazour@ucf.edu

[2] Debjit Roy, Indian Institute of Management Ahmedabad, Gujarat-380015, India, Email: debjit@iimahd.ernet.in

travel requirements may walk-in at a depot to rent a vehicle without a reservation. Therefore, customers can be broadly classified into two classes: reserve and walk-in.

Reserve customers place a vehicle reservation either directly with a rental provider or through a third-party travel site. Reservations provide a benefit to customers by increasing the chance that a vehicle will be available when the customer arrives and to vehicle providers through increased sales and for planning purposes. However, reservations require holding idle vehicles, which can reduce fleet utilizations especially when customers' arrival times vary from the information provided in their reservation (either by not arriving to pick up their vehicle or by arriving late).

Walk-in customers arrive to the rental depot without a prior reservation to request a rental vehicle. Walk-in customers may not get a vehicle immediately on arrival and may wait longer for a vehicle. Walk-in customers are common at local market depots where repair or insurance replacements make up a significant portion of sales. Rental providers need to serve both types of customers to sustain growth and improve fleet utilization [6]. Walk-in customers represent an important revenue stream; consequently, the walk-in customer experience must also be managed effectively.

Both types of customers are served by the same fixed fleet of vehicles. This fixed fleet is a major expense for rental providers as the vehicles' values quickly depreciate with time. Hence, a key to profitability is to maintain high vehicle utilization, which can be achieved by sustained sales. Sustainable sales are achieved by providing high quality customer service. In the rental vehicle industry, overall customer satisfaction is determined by both responsiveness and customer differentiation. Responsiveness is often measured by a customer's length of stay at a depot [6, 23]. Customer length of stay is composed of waiting times due to vehicle unavailabilities, waiting time to see a sales representative, and service times. We focus on waiting times due to vehicle unavailabilities. A discount is typically offered to customers who have to wait for a vehicle to become available. We account for this loss of revenue, as well as loss of customer goodwill in our problem with a penalty for customer waiting times. Also, we focus on local market depots that are located away from competitor locations (i.e., non-airport locations). The local market depots represent a significant portion of rental provider revenues and there are numerous such depots in metropolitan areas. For example, local markets represented approximately half of Enterprise Rent-a-Car's 2010 revenues [5]. At these depots, switching among providers is not convenient and customers typically wait for a rental vehicle instead of balking.

This research also accounts for several operational uncertainties in the vehicle allocation process such as demand uncertainties and vehicle availabilities. Demand forecasts form the basis for decision making in the rental vehicle industry and errors in forecasts are derived from the high rate of no-shows, last-minute bookings, high rate of walk-in customers, and the uncertainty of rental length [7, 30]. For example, it was recently reported that 41% of rental vehicle bookings made through priceline.com's mobile apps were same-day and among these bookings 48% were made within two hours of pickup [22]. Also, rental providers can experience non-stationary demand patterns. The rental length varies among customers, with a high likelihood of customers renting a vehicle for a few days and lower likelihood of customers renting a vehicle for longer time periods. A high degree of uncertainty of when a vehicle is available also exists. Such uncertainty is caused by delayed check-ins, variability and congestion in the inspection, refueling, and cleaning process, damaged returned vehicles that require time for repair, and inaccurate inventory status in the company's information system [7]. We define the vehicle unavailability period as the time that lapses from when a customer is assigned to the vehicle until the vehicle is available to be rented to a new customer. The vehicle unavailability period includes the time the customer is rented the vehicle, the time for check-in, inspection, and cleaning, and the time the vehicle is out due to maintenance of a damaged vehicle. Because of these uncertainties, a common practice for rental vehicle companies is to flexibly handle the allocation of vehicles to customers when the customer arrives [7].

Reserve customers receive priority over walk-in customers during the vehicle assignment process; however, there is uncertainty associated with when the reserve customer will arrive and when a vehicle will become available. Therefore, providers face the following allocation dilemma, "should an available vehicle be allocated to a walk-in customer given customer reservations exist?" Customer service representatives must answer this question each time a walk-in customer arrives, realizing there is a trade-off between accepting the walk-in customer and obtaining the associated profit or having the walk-in customer wait to keep available vehicles free for reserve customers. If the vehicle is allocated to the walk-in customer, the waiting time of the walk-in customer decreases and the utilization of the vehicle improves at the expense of possibly increasing the waiting time of the reserve class customers. On the other hand, if the available vehicle is not allocated to the walk-in customer, the vehicle is not utilized and the waiting time of the walk-in customer increases. Walk-in customers are more endurable to waiting times than reserve class customers; therefore, holding vehicles for reserve customers adds value particularly because the waiting costs for reserve class customers are greater than for the waiting costs for walk-in customers.

The two types of customers, as well as customer demand and vehicle supply uncertainties, make determining how best to allocate capacity a difficult operational decision. Currently, rental providers allocate capacity dynamically, providing a walk-in customer a vehicle depending on the number of vehicles available. Hence identifying an optimal threshold quantity (denoted as K^*) for vehicle allocation to the walk-in class is of significant interest to rental providers. Vehicles will be assigned to the walk-in customers only if the number of vehicles on hand exceeds K^*. Such a policy provides flexibility because the decision is contingent on the current state of the system and determining whether to provide walk-in customers with a vehicle or to have them wait for an available vehicle is not set a priori.

Our contribution lies in developing a stochastic model that can handle the uncertainty of demand arrivals and vehicle unavailabilities based on the state of the system to determine under what conditions to allocate vehicles to walk-in customers that considers the expected waiting time of both reserve and walk-in customers. Previous works analyze the vehicle rental system using a customer loss model (for example, see [8, 24]). Thus, this research is an initial attempt to develop a rental depot profitability model by considering customer waiting times for multiple classes. Each depot is modeled with a multi-class non-work-conserving semi-open queue with stochastic inputs, where the vehicles are not allocated to walk-in customers beyond a threshold quantity. We provide an optimization formulation to determine the optimal threshold quantity, K^*. We use analytical models to solve the optimization formulation when arrival rates are stationary for both exponential and deterministic vehicle unavailability distributions. We develop threshold policies for non-stationary arrival patterns and test their performance with a discrete-event simulation model. Through analysis and testing, we provide insights into the operating characteristics that impact the threshold quantities and present counterintuitive results.

The remainder of this paper is organized as follows. In Section 2, we review relevant literature on priority customer class-based queuing systems. In Section 3, we provide our queuing network model, which includes a continuous-time Markov chain representation of the rental system network and our optimization formulation. In Section 4 we analyze the behavior of the optimal threshold quantity for stationary arrival rates. We then extend our work to include threshold policies for non-stationary arrival rates in Section 5. Finally, in Section 6, we provide conclusions and future research directions.

2. Literature Review

Priority customer class-based queuing systems are commonly seen among a spectrum of service industries where both responsiveness and customer differentiation determine the overall customer satisfaction. Examples include communication networks [15, 16], manufacturing and inventory systems [12, 28], transportation networks [8, 20, 21, 24], and healthcare applications [13, 26, 27, 29]. Research has been conducted on "admission policies" for multiple classes of customers, which determine whether a company should accept an arrival from a specific class of customers or not. In this line of research, if the customer is not offered service, the customer typically balks and leaves the system (i.e., the network is modeled as a loss system). In this section, we first review analytical models for class-based loss system models, emphasizing the use in rental systems, and then review literature on threshold policies for multi-class priority queuing systems.

Class-Based Loss System Models: Savin et al. [24] develop a multiple-server loss system of a rental problem where capacity is rationed among two customer classes. In addition to determining an admission policy, they consider how the use of tactical controls affects longer-term fleet-sizing decisions and develop an aggregate threshold heuristic based on a fluid approximation of the original stochastic model. Gans and Savin [8] develop a queuing loss model for rental businesses with two customer classes: contract customers and walk-in customers. They derive an optimal admission policy for contract customers and optimal prices for walk-in customers, deciding when to offer service to contract customers and what fees to charge walk-in customers for service. Papier and Thonemann [20] determine an admission policy for two customer classes in the rail cargo industry, premium and classic service, where advance demand reservation information about the premium class is available. Papier and Thonemann [21] develop an algorithm that simultaneously determines the fleet size and the admission policy in the rail cargo industry with batch arrivals rates by developing a batch-arrival queuing loss system. George and Xia [9] develop a closed queuing network for optimal fleet-sizing decisions of a vehicle-sharing rental provider.

The above literature examines admission policies that determine when to deny service to lower priority customers. In such a situation, queuing loss models are appropriate and the expected waiting time of customers is not considered. Our work varies in that we are considering policies where walk-in customers do not balk, and instead wait until the number of vehicles available is greater than a threshold quantity. In such systems, the primary performance metric is to minimize the cost of waiting for both customer classes. Because reserve customers have a higher waiting penalty, a priority threshold queuing system can be used.

Priority threshold queuing systems: Miller [17] develops a Continuous Time Markov Chain (CTMC) for an n-server queuing system with m customer classes and proves the optimality of threshold policies; however, does not determine optimal policy parameters. Altman et al. [1] apply a dynamic programming methodology to determine optimal capacity allocation rules when multiple customer classes use shared resources with no waiting space. Ormeci et al. [18] also uses dynamic programming to study a similar situation with two customer classes and no waiting room.

Taylor and Templeton [29] study a multi-server cutoff-priority queue for determining the number of ambulances required to transfer both emergency and lower priority patients. Using probability generating functions, they obtain explicit expressions for: 1) the probability of n servers being busy, 2) the Laplace-Stieltjes transform of the low priority waiting time distribution, 3) the expected low priority waiting time, and 4) the complete low priority waiting time distribution (the inversion of the Laplace-Stieltjes transform). They determine a threshold quantity for the number of beds reserved for future emergency customer arrivals. These results are applied to determine the number of ambulances required in an urban fleet that serves both emergency calls and low priority patient transfers. Their work is extended in [25] to consider more than two classes of customers.

The work in [29] most closely aligns with our work to develop threshold policies for a vehicle rental system with two customer classes; however, our work can be differentiated in the following ways:

- We propose a queuing network model based on non-work-conserving semi-open queues that permits the analysis of the performance of any topology of vehicle rental networks. Further, the solution approach permits determining the distribution of vehicles at all stations.
- We incorporate the queuing model in to an optimization framework and analyze the optimal threshold quantity that minimizes weighted waiting costs. In doing so, we provide valuable, counterintuitive results.
- We consider exponential and deterministic service times and analyze multiple threshold policies with both stationary and non-stationary arrivals.

Next, we describe the basic model of a vehicle rental depot with two customer classes.

3. Queuing Network Model

To model a vehicle rental depot, we initially assume that the arrival process for both types of customers (reserve and walk-ins) is Poisson with stationary arrival rates and the vehicle unavailability period follows the exponential distribution. We analyze systems with non-stationary arrivals in Section 5 and deterministic vehicle unavailability periods in Section 4.1. Further, we assume customers wait for a vehicle to be available and do not abandon the queue.

3.1 Performance Objective of the Rental Provider

We describe the performance objective of the rental provider using a nonlinear optimization formulation. The formulation is developed to minimize a weighted combination of customer waiting costs and the decision variable is the threshold quantity of vehicles that are held in anticipation for the reserve customers.

We let the set of customer classes be indexed by i, where $i \in \{r \text{ (reserve)}, w \text{ (walk-in)}\}$. We let e_i denote the waiting penalty of customer class i, k the threshold quantity, and $\mathbb{E}[W_i(k)]$ the expected waiting time per customer of class i, which is a function of the threshold quantity k.

The objective function, denoted as Θ_k and expressed by Equation 1, is the expected waiting costs of the two customer classes weighted by waiting penalties. This *Weighted Waiting Cost* objective incorporates the trade-off between walk-in and reserve customer waiting costs and is a function of the decision variable k. The objective function is nonlinear due to the $\mathbb{E}[W_i(k)]$ terms. As the threshold quantity increases, the expected reserve customer waiting time decreases, while the expected walk-in customer waiting time increases. The constraints include that the system is stable (i.e., the system utilization ρ is less than one), and that the threshold quantity be a non-negative integer and less than the number of vehicles, V. We denote the optimal threshold quantity that minimizes Equations 1 - 4 as K^*.

$$\begin{aligned} \min_k \Theta_k &= \sum_{i=\{r,w\}} e_i \mathbb{E}[W_i(k)] && (1) \\ \text{s.t.} \quad & \rho < 1 && (2) \\ & 0 \leq k \leq V && (3) \\ & k \in \{\mathbb{Z}^+\} && (4) \end{aligned}$$

For simplicity, we refer to $\mathbb{E}[W_i(k)]$ as $\mathbb{E}[W_i^k]$ in the remainder of the paper. To solve the optimization model, we use a queuing model to obtain the expected waiting time per customer of both classes, which is the focus of the next section.

3.2 A Queuing Model with Two Customer Classes to Obtain the Expected Waiting Times

As displayed in Figure 1a, we develop a queuing model with two customer classes. The interarrival times of the walk-in and reserve customers to the depot are IID exponential random variables with means $(\lambda_w)^{-1}$ and $(\lambda_r)^{-1}$ respectively. Customers on arrival to the depot wait at buffer B_1. The depot has a fixed fleet size of V vehicles. Vehicles wait for allocation at buffer B_2. When a reserve class customer arrives, the customer is matched with a vehicle at the synchronization station J. However, when a walk-in customer arrives, the customer is matched with a vehicle only if the number of available vehicles is more than the threshold quantity k. A threshold policy facilitates that reserve customers receive priority access to vehicles over the walk-in customers. Therefore, vehicles and walk-in customers can wait simultaneously at their respective buffers, which results in a non-work-conserving queue. Service is first-come-first-served within each class of customers. Once a customer and vehicle are matched, the vehicle is unavailable to rent for an exponential amount of time with mean μ^{-1}. The vehicle unavailability period is modeled as an Infinite Server (IS) station. We denote this network as a Non-work-conserving Infinite Server Semi-open Queuing Network (NIS-SOQN). We can evaluate the network using a CTMC with a two-tuple state vector (y, i), where y is the difference between the number of reserve class customers waiting in buffer B_1 and the number of vehicles available in buffer B_2, $y \in \{-V, \dots, \infty\}$. The second component i denotes the number of walk-in class customers waiting for a vehicle, $i \in \{0, \dots, \infty\}$. The transitions among states are provided in Table 1.

Note that a CTMC provides a numerical solution-based approach that is adaptable to complex network topologies. For instance, upon return, vehicles may require maintenance. This topology can be modeled using a queuing network model as displayed in Figure 2, where we separate the vehicle unavailability period into two components. The vehicle rental periods are modeled as IID exponential random variable with mean μ_r^{-1} and the maintenance times are model as IID exponential random variable with mean μ_m^{-1}. Note that while most of the customers are expected to rent a vehicle for a few days (2 or 3 days), there are customers who rent vehicles for a longer period (for instance,

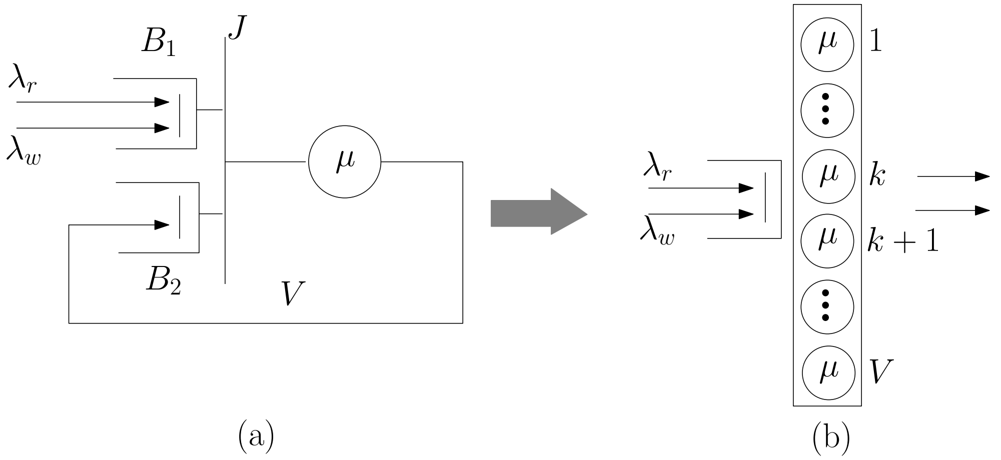


Figure 1: (a) Semi-open queuing model of the rental system with two customer classes, (b) an equivalent $M/M/V$ queue with two customer class and cutoff priority queue

Table 1: Transitions in the CTMC for SOQN with multiple customer classes

Condition	**Rate**	X_j
$y = -V$	$\lambda_w + \lambda_r$	$(y+1, i)$
$-V < y < -T$	$\lambda_w + \lambda_r$ $(V+y)\mu$	$(y+1, i)$ $(y-1, i)$
$y = -T, i = 0$	λ_r λ_w $(V+y)\mu$	$(y+1, i)$ $(y, i+1)$ $(y-1, i)$
$y = -T, i > 0$	λ_r λ_w $(V+y)\mu$	$(y+1, i)$ $(y, i+1)$ $(y, i-1)$
$-T < y \leq 0, i \geq 0$	λ_r λ_w $(V+y)\mu$	$(y+1, i)$ $(y, i+1)$ $(y-1, i)$
$y > 0, i \geq 0$	λ_r λ_w $(V)\mu$	$(y+1, i)$ $(y, i+1)$ $(y-1, i)$

as a replacement vehicle). Hence, an exponential distribution would capture the large variation in rental times. If the probability that a vehicle requires maintenance is p_m, the vehicle is routed to the maintenance station with probability p_m and returned to the idle pool of vehicles in buffer B_1 with probability $1 - p_m$. The definition of the state variable in the CTMC is expanded by adding another component, $j : j = 1, \ldots, V$, to account for the number of vehicles undergoing maintenance. For numerical experimentation, we consider $p_m = 0.1$, $\mu_r^{-1} = 2$ days, $\mu_m^{-1} = 1$ day, and $V = 20$.

Figure 3 presents the weighted objective function value for different threshold quantities and different ratios of waiting time penalties (i.e., $\frac{e_r}{e_w}$). We can see that when the penalty ratio is less than 10, $K^* = 1$. However, when the ratio is 50, more vehicles are to be held in anticipation of reserve customer arrivals and $K^* = 2$.

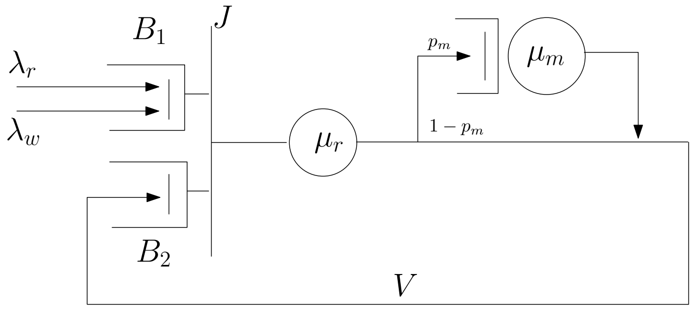


Figure 2: Queuing model of the rental system with maintenance function

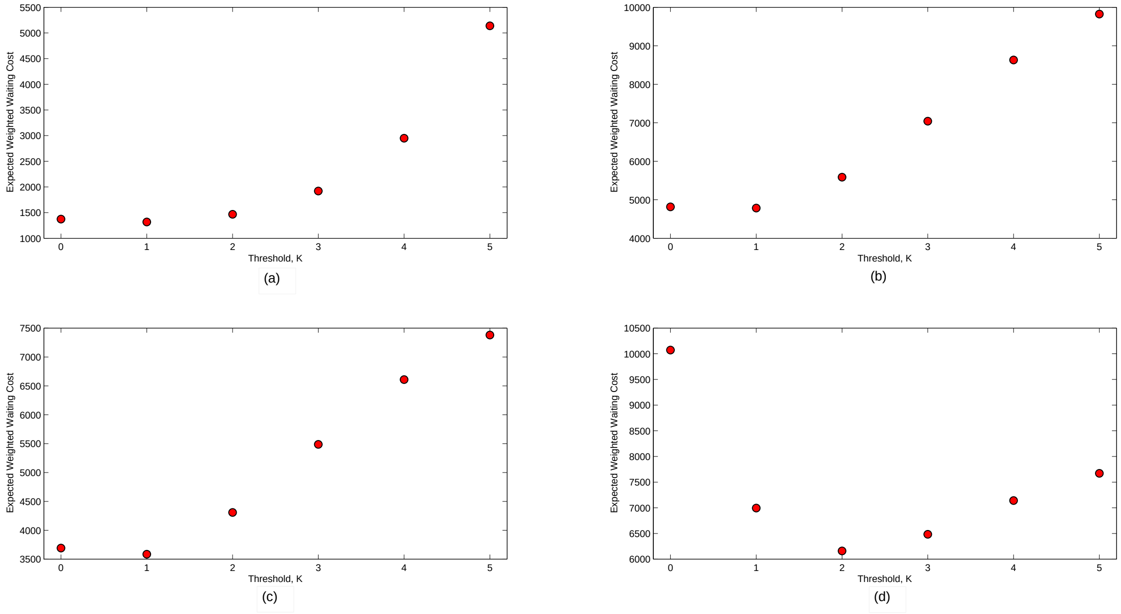


Figure 3: The behavior of the expected weighted waiting costs for different threshold quantities for (a) $\lambda_w = 2$, $\lambda_r = 5$, $\frac{e_r}{e_w} = 10$ (b) $\lambda_w = 3$, $\lambda_r = 5$, $\frac{e_r}{e_w} = 10$ (c)$\lambda_w = 4$, $\lambda_r = 4$, $\frac{e_r}{e_w} = 10$ (d) $\lambda_w = 4$, $\lambda_r = 4$, $\frac{e_r}{e_w} = 50$, V=20

The state-space of a CTMC grows exponentially with the number of components and the range of each component associated with a state vector. Due to this state dimensionality issue, it is computationally expensive to evaluate the system performance when the number of vehicles in the system increases. Therefore, we develop an equivalent queuing model of the work-conserving SOQN with two classes and one IS queue with exponential service times that is computationally efficient to evaluate (see Figure 1b).

3.2.1 Equivalent Model of the SOQN with Priority Customers and one IS Station.

In this section, we propose a *multi-server cutoff priority queue*, which is equivalent to the SOQN with an IS station and priority class customers (NIS-SOQN). Taylor and Templeton [29] considered a priority queue in steady state with V servers, two classes of customers, and a threshold allocation policy. In their model, lower priority customer arrivals are refused immediate service and placed in a queue whenever $V - k$ or more servers are busy, in order to keep k servers free for high priority arrivals. We now present the theorem, that formally states the model equivalency.

Theorem 1 *For a NIS-SOQN with V vehicles, two customer classes, exponential inter-arrival times, and one IS station with an exponential service time distribution, the multi-server cutoff priority queue (where each vehicle corresponds to a server in the multi-server queue) proposed in [29] provides an exact estimate of the queue length distribution and the throughput at the external queue for both customer classes.*

proof: The equivalence can be shown by comparing the states of the CTMC for NIS-SOQN and the multi-server cutoff priority queue (see Figure 1b). We can also evaluate the multi-server cutoff priority queue using a CTMC with a two-tuple state vector (y, i), where y is the difference between the number of reserve class customers waiting in buffer and the number of idle servers, $y \in \{-V, \dots, \infty\}$. The second component i denotes the number of walk-in class customers waiting for a vehicle, $i \in \{0, \dots, \infty\}$. The transitions among the states are also similar to the ones included in Table 1; therefore, the models are equivalent.

We use the exact expression in [29] for the expected waiting time of walk-in (low priority) customers when k vehicles are held for reserve (priority) customers. The expected waiting time for walk-in customers with k vehicles reserved, $\mathbb{E}[W_w^k]$, is displayed in (5), which has been adapted to reflect the notation in this paper.

$$\mathbb{E}[W_w^k] = P_0\mu^{-1}\left(\frac{\nu_w + \nu_r}{\nu_r}\right)^{V-k}(V-k)^2\left[(V-k) - \nu_w S(V-k,V)\right]^{-2}\left[\frac{\nu_r^V V^2}{V!(V-\nu_r)^3} + \sum_{i=V-k}^{V-1}\frac{\nu_r^i}{i!}\frac{(S(i,V))^2}{i}\right] \tag{5}$$

where $\nu_w = \frac{\lambda_w}{\mu}$; $\nu_r = \frac{\lambda_r}{\mu}$; $P_0 = \left[\left(\frac{(\nu_r+\nu_w)^{V-k}}{(V-k-1)!}\right)\left(\frac{S(V-k,V)}{V-k-\nu_w S(V-k,V)}\right) + \sum_{i=0}^{V-k-1}\frac{(\nu_r+\nu_w)^i}{i!}\right]^{-1}$;

and $S(j,V) = \nu_r^{-j} j!\left[\frac{\nu_r^V V}{V!(V-\nu_r)} + \sum_{i=j}^{V-1}\frac{\nu_r^i}{i!}\right]$.

As the threshold priority queue assumes exponential service times, the high priority waiting time W_r is exactly as for the $M/M/V$ queue except for the change in the probability that all vehicles are busy, P_V. From [29], $P[W_r > t] = P_V e^{-(V\mu-\lambda_r)t}$. Therefore, the expected waiting time for the reserve customers with a threshold quantity k, $\mathbb{E}[W_r^k]$, can be derived from the probability distribution function in Equation (6), and is displayed in Equation (7).

$$f_{W_r}(t) = P_V(V\mu - \lambda_r)e^{-(V\mu-\lambda_r)t}; t \geq 0 \tag{6}$$

$$\mathbb{E}[W_r^k] = \frac{P_V}{(V\mu - \lambda_r)} \tag{7}$$

where $P_V = P_0 \left(\frac{\rho^{V-k}\nu_r^k}{V!}\right)\left(\frac{V-k}{V-k-\nu_w S(V-k,V)}\right)\left(\frac{V}{V-\nu_r}\right)$ and $\rho = \frac{\lambda_r+\lambda_w}{\mu}$.

Next, we explore how different threshold quantities impact our objective of minimizing weighted waiting costs.

4. Optimal Threshold Quantities with Stationary Arrivals

In this section we analyze how the optimal threshold quantity changes for different reserve and walk-in customer arrival rates. We apply the closed-form solutions for the expected waiting time for walk-in and reserve customers as defined in Equations 5 and 7, respectively. To determine the optimal threshold quantity (denoted as K^*), we enumerate the expected waiting times for all integer values between 0 and V, and select the optimal threshold quantity as the k-value that minimizes the nonlinear model in Equations 1 - 4.

In Figure 4, we illustrate the behavior of the objective function for different threshold quantities and reserve customer arrival rates. The x-axis denotes the threshold quantity and the y-axis the objective value of expected weighted waiting costs, where $\mathbb{E}[W_i^k]$ is in minutes. The number of vehicles available is 25 and the mean vehicle unavailability is 2 days. Figure 4 illustrates the nonlinear behavior of the objective function. The optimal threshold quantity is the value that minimizes the weighted waiting costs and varies depending on the reserve customer arrival rate. In Figure 4, $K^* \leq 3$ regardless of the value of λ_r.

Figure 5 presents the optimal threshold quantity as a function of the reserve customer arrival rate for different number of vehicles, walk-in rates and penalty ratios. Interestingly, the optimal threshold quantity does not always increase as the reserve customer arrival rate increases. Instead, K^* is a concave function that increases (up to a point) with an increase in reserve customer arrivals

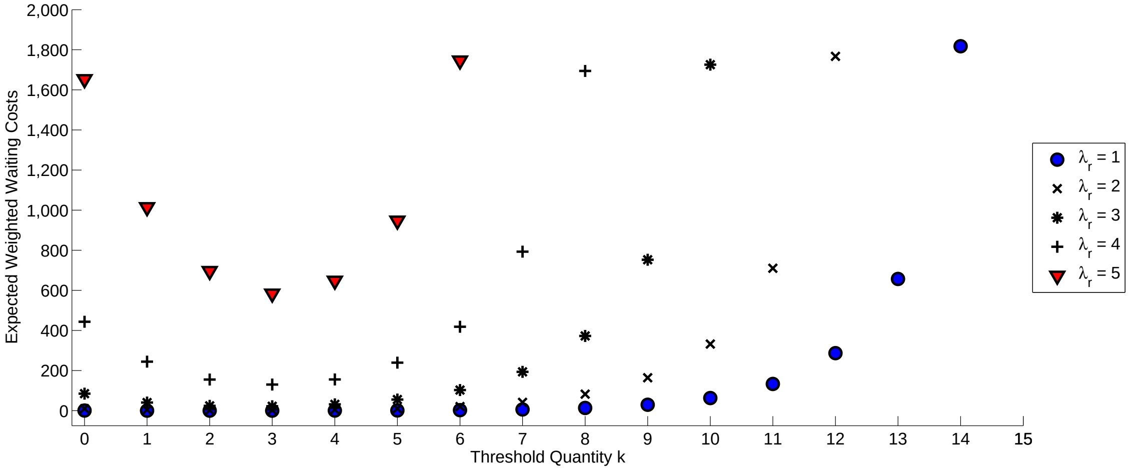


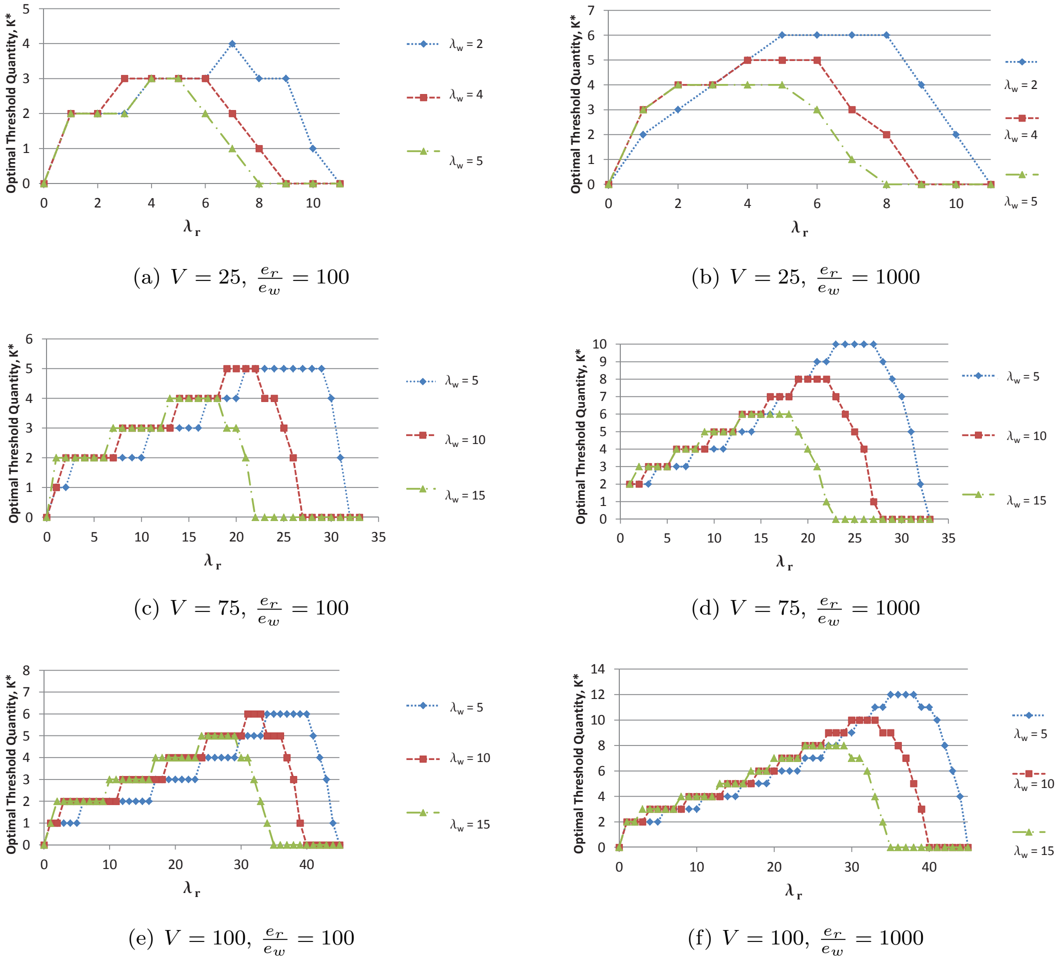
Figure 4: The weighted waiting costs as a function of the threshold quantity for a system with $V = 25$, $\mu = 0.5$, $\lambda_w = 4$, $e_r = 100$, and $e_w = 1$.

and then decreases. For instance, in Figure 5a with $\lambda_w = 2$, K^* increases upto 4 when $\lambda_r = 7$. As λ_r increases beyond 7 per day, K^* decreases.

In addition, the optimal threshold value does not always increase as the walk-in arrival rate increases. Instead, the threshold value increases as the walk-in arrival rate increases up to a point and then decreases. For example, in Figure 5(c), K^* with $\lambda_w = 5$ is less than or equal to K^* with $\lambda_w = 15$ if $0 \leq \lambda_r \leq 18$. However, if $\lambda_r \geq 18$, the reverse occurs and K^* with $\lambda_w = 5$ is greater than or equal to K^* with $\lambda_w = 15$.

The reason for these two insights is that increasing the reserve customer arrival rate results in higher utilization of the vehicles in the system. The results from classical queuing theory indicate that the expected customer waiting time in a queue increases nonlinearly with an increase in server utilization. Waiting times are very sensitive to utilization, especially at high utilization values. This sensitivity of system performance to utilization results in setting a threshold quantity that can actually decrease as reserve customer arrival rates increase. Also, the threshold quantity can decrease as the walk-in arrival rate increases. As threshold policies are set to ensure that reserve customer waiting times are reduced, these can be a counterintuitive results.

As displayed in Figure 5, the optimal threshold value varies depending on the penalty ratio (i.e., e_r/e_w). As the weight placed on reserve customers increases, the number of vehicles reserved increases, which results in larger threshold quantities being optimal. When reserve and walk-in



(a) $V = 25$, $\frac{e_r}{e_w} = 100$

(b) $V = 25$, $\frac{e_r}{e_w} = 1000$

(c) $V = 75$, $\frac{e_r}{e_w} = 100$

(d) $V = 75$, $\frac{e_r}{e_w} = 1000$

(e) $V = 100$, $\frac{e_r}{e_w} = 100$

(f) $V = 100$, $\frac{e_r}{e_w} = 1000$

Figure 5: Optimal K^* with $\mu = 0.5$ and (a) $V = 25$, $\frac{e_r}{e_w} = 100$, (b) $V = 25$, $\frac{e_r}{e_w} = 1000$, (c) $V = 75$, $\frac{e_r}{e_w} = 100$, (d) $V = 75$, $\frac{e_r}{e_w} = 1000$, (e) $V = 100$, $\frac{e_r}{e_w} = 100$, and (f) $V = 100$, $\frac{e_r}{e_w} = 1000$.

customer waiting times are given equal weights (i.e., $e_r/e_w = 1$), the optimal threshold policy is always to set $K^* = 0$ (see Lemma 1). This results in the highest utilization of a fixed fleet of vehicles.

Lemma 1 *If $\frac{e_r}{e_w} = 1$, i.e., reserve and walk-in customer waiting times are given equal weights then $K^* = 0$.*

proof: Since $e_r = e_w = e$, the objective value in Equation 1, $\Theta_k = e \sum_{i=\{r,w\}} \mathbb{E}[W_i(k)]$. Let U_{k_0} and $U_{k_{>0}}$ be the utilization of the system with $K^* = 0$ and $K^* > 0$, respectively. The utilizations are related as follows: $U_{k_{>0}} > U_{k_0}$. Therefore, $\mathbb{E}[W_i(k > 0)] > \mathbb{E}[W_i(k = 0)]$. Hence, $K^* = 0$ is optimal.

4.1 Threshold Policies with Deterministic Service Times

Exponential service times model systems with high variability in the vehicle unavailability period. Some depots may experience low variability in the vehicle unavailability period; therefore, we analyze the threshold policy of such systems by developing an approximate model for a deterministic service time distribution.

To model the case with a constant, deterministic vehicle unavailability period, we develop an approximation formula. We use the results of an $M/D/V$ queue with two nonpreemptive priority classes (see [3]). Note in [3], the authors do not consider threshold policies. To approximate the expected waiting time for both classes with deterministic service times, the methodology in [3] is to multiply the expected waiting time with exponential service times by correction factors. The correction factor for the reserve class is $\left(\frac{(1-f\rho)V}{V+1}+\frac{f\rho}{2}\right)$ and the correction factor for the walk-in class is $\left(\frac{(1-(1-f)\rho)V}{V+1}+\frac{(1-f)\rho}{2}\right)$. The percentage of reserve customers is denoted by $f=\lambda_r/(\lambda_r+\lambda_w)$.

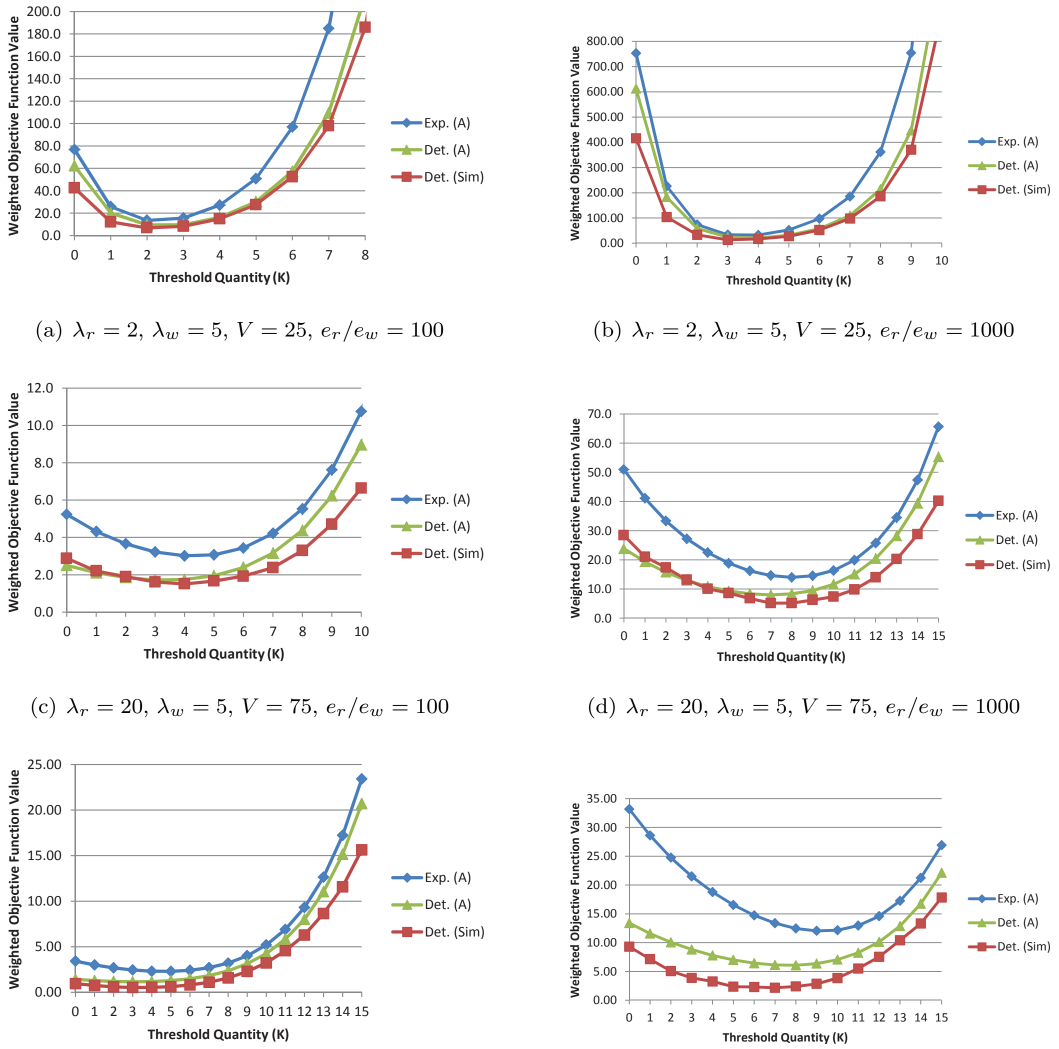
To approximate the average waiting time with a deterministic vehicle unavailability period and *a threshold policy* for two nonpreemptive priority classes, we adjust the expected waiting time with exponential periods, $\mathbb{E}[W_r^k]$ and $\mathbb{E}[W_w^k]$, with the factors from [3]. The approximations for the expected waiting time with deterministic periods for reserve and walk-in customers, denoted as $\mathbb{E}[W_r^D]$ and $\mathbb{E}[W_w^D]$, are presented in Equations 8 and 9, respectively,.

$$\mathbb{E}[W_r^D]=\mathbb{E}[W_r^k]\left(\frac{(1-f\rho)V}{V+1}+\frac{f\rho}{2}\right) \tag{8}$$

$$\mathbb{E}[W_w^D]=\mathbb{E}[W_w^k]\left(\frac{(1-(1-f)\rho)V}{V+1}+\frac{(1-f)\rho}{2}\right) \tag{9}$$

We compare the approximate deterministic expressions against a discrete-event simulation model and the exponential expressions, denoted as *Det. (A), Det. (Sim),* and *Exp. (A)*, respectively. The expected waiting times for *Det. (A)* are calculated using Equations 8 and 9; for *Exp. (A)* from Equations 5 and 7. The expected waiting time for *Det. (Sim)* are determined from the results of 10 replications of 2,500 days with a 500 day warm-up period. Figure 6 displays the weighted objective function for the case of $\mu = 0.5$ for different reserve and walk-in arrival rates, number of vehicles, and penalty weights. As compared to the simulation model, the analytical model for deterministic service time distribution does a good job at approximating the objective function. For Figures

6(a), 6(c), and 6(d), the threshold quantity that minimizes the objective function value is the same for *Det. (Sim)* and *Exp. (A)*. In the remaining cases, the weighted objective function cost is fairly insensitive to the threshold quantity.



(a) $\lambda_r = 2$, $\lambda_w = 5$, $V = 25$, $e_r/e_w = 100$

(b) $\lambda_r = 2$, $\lambda_w = 5$, $V = 25$, $e_r/e_w = 1000$

(c) $\lambda_r = 20$, $\lambda_w = 5$, $V = 75$, $e_r/e_w = 100$

(d) $\lambda_r = 20$, $\lambda_w = 5$, $V = 75$, $e_r/e_w = 1000$

(e) $\lambda_r = 30$, $\lambda_w = 5$, $V = 100$, $e_r/e_w = 100$

(f) $\lambda_r = 30$, $\lambda_w = 5$, $V = 100$, $e_r/e_w = 1000$

Figure 6: Comparsion of the approximate deterministic expressions (*Det. (A)*) against a discrete-event simulation model (*Det. (Sim)*), and the exponential expressions (*Exp. (A)*) for varying arrival rates, number of vehicles, and penalty ratios.

In the next section we consider rental systems with time-varying arrival rates for customers.

5. Threshold Policies for Non-Stationary Demand

Threshold policies are designed to provide a differentiated service to reserve customers and reserve customer arrival rates can be non-stationary. Therefore, we analyze the impact that non-stationary reserve arrival patterns have on threshold policies. To model such a system, we let $\lambda_r(t)$ denote the reserve customer arrival rate at time t. We assume arrival patterns repeat over a planning period of length T and assume stationary arrival rates for walk-in customers. As a way to evaluate the threshold policies, we use the objective function in Equation 1. We denote the average daily reserve and walk-in customer arrival rate for the planning period as $\overline{\lambda_r}$ and $\overline{\lambda_w}$, respectively. We define the fleet utilization, $\rho = \frac{(\overline{\lambda_r}+\overline{\lambda_w})}{V\mu}$.

We analyze the following seven threshold policies for non-stationary demand.

1. **Average Policy.** In this policy, a constant threshold quantity is set by applying the optimal threshold quantity for the average stationary arrival rate over the entire planning period (i.e., we use $\overline{\lambda_r}$ and $\overline{\lambda_w}$ to find K^* as in Section 4).

2. **Stationary Independent Period by Period (SIPP) Policy.** In this policy, a dynamic threshold quantity is set by using the optimal threshold quantity assuming the system can be divided into independent stationary periods [11]. We let $k^*(t)$ denote the threshold policy at time t and is calculated using $\lambda_r(t)$ and λ_w as in Section 4. This is a dynamic policy, because as the arrival rate changes, the threshold quantity also changes.

3. **Pointwise Stationary Approximation (PSA) Policy.** In this policy, a constant threshold quantity is set by computing the average time-weighted threshold quantity over the planning period using the optimal stationary threshold that correspond to each point in time [10]. More formally, $K^* = (\sum_t k^*(t))/T$, where $k^*(t)$ is obtained from the SIPP Policy.

4. **Average Stationary Approximation (ASA) Policy.** In this policy, a constant threshold policy is set by calculating the stationary threshold quantity for each arrival rate and then computing the reserve demand-weighted threshold average over the planning period [19]. More formally, $K^* = (\sum_t(\lambda_r(t)k^*(t))/(\sum_t \lambda_r(t))$, where $k^*(t)$ is obtained from the SIPP Policy.

5. **Maximum Threshold Policy.** In this policy, a constant threshold quantity is set by using the maximum threshold quantity found via the SIPP policy. More formally, $K^* = \max_t\{k^*(t)\}$, where $k^*(t)$ is obtained from the SIPP Policy.

6. **Minimum Threshold Policy.** In this policy, a constant threshold quantity is set by using the minimum threshold quantity found via the SIPP policy. More formally, $K^* = \min_t\{k^*(t)\}$, where $k^*(t)$ is obtained from the SIPP Policy.

7. **Priority Without a Threshold Policy.** In this policy, no vehicles are held for reserve customers (i.e., $K^* = 0$). However, reserve customers are still given non-preemptive priority over walk-in customers.

For all policies, we require K^* to be integer. If a policy results in a non-integer value, we round the threshold quantity to the nearest integer value.

5.1 Numerical Experiments

Because we are aware of no tractable analytical model that can evaluate the multi-server multi-class priority queue with a threshold policy for non-stationary demands, we develop a discrete-event simulation model to evaluate the identified threshold policies. The discrete-event simulation model of the vehicle rental process was constructed using Arena Simulation Software (*www.arenasimulation.com*). As indicated in Figure 7, there are two modules in the model: the main rental depot model and a threshold quantity control model.

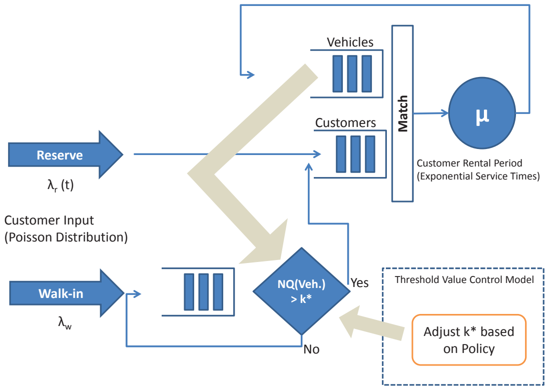


Figure 7: Discrete-Event Simulation Model of a Rental System with a Threshold Policy

The data for our numerical experiments are based on discussions with a large rental provider, a case study [5], and academic and trade literature [4, 14, 6]. We assume reserve customers have a non-stationary arrival pattern with exponential inter-arrival times, but walk-in arrivals have a stationary pattern with exponential inter-arrival times. We define $p(t)$ as the daily expected demand of day

t as a percentage of the total number of reserve customers expected to arrive in one week. For our analysis, we use a trianglular pattern and set $p(1) = p(7) = 0.0625$, $p(2) = p(6) = 0.125$, $p(3) = p(5) = 0.1875$, and $p(4) = 0.250$. For example, on day 4, 25% of the weekly demand occurs. If $\overline{\lambda_r} = 15$, on day 4, 26.25 reserve customers are expected to arrive (i.e., $(0.25)(15)(7) = 26.25$).

We conduct a full factorial experimental design with the following factors, resulting in 36 scenarios:

- Number of vehicles, $V = 25, 75, 100$.
- Fleet utilization, $\rho = 0.40, 0.60, 0.80$.
- Relative number of reserve customers to walk-in customers, $\overline{\lambda_r} = \overline{\lambda_w}$ and $\overline{\lambda_r} = 4\overline{\lambda_w}$.
- Penalty ratio, $e_r/e_w = 100, 1000$.

By varying the fleet utilization and the relative percentage of reserve customer and walk-in customers, we test different $\overline{\lambda_r}$ and $\overline{\lambda_w}$ values. The vehicle unavailability periods are assumed to follow an exponential distribution with a mean of 2 days.

Table 2 provides values for the threshold policies for each scenario. In many scenarios, different threshold policies result in the same K^* values (e.g., Policies 1, 3, 4, and 6 all result in $K^* = 2$ for scenario 1). Because we have a symmetric arrival pattern, we denote the SIPP policy with a four-tuple vector, $(k(1) = k(5), k(2) = k(6), k(3) = k(7), k(4))$.

For each threshold policy and scenario, Tables 3 - 6 present system performance metrics. Simulation results are presented for the average of 10 replications of length 2500 days with a 500 day warm-up period. The performance metrics include the expected waiting time (in minutes) for reserve and walk-in customers, as well as the weighted waiting cost objective function, θ_k. For each scenario, the policies are presented by decreasing objective function values and the minimum objective function value is presented in bold font.

When the best threshold policy is applied, the average reserve and walk-in customer waiting time over all scenarios are 10.11 and 295.96 minutes, respectively. Therefore, a threshold policy allows for reserve customer waiting times to be manageable at the expense of higher walk-in customer waiting times. As expected, the average waiting time varies based on fleet utilization (average reserve customer waiting times are 0.01 minutes for $\rho = 0.40$, 1.07 minutes for $\rho = 0.60$, and 29.24 minutes for $\rho = 0.80$; average walk-in customer waiting times are 1.50 minutes for $\rho = 0.40$, 56.26 minutes for $\rho = 0.60$, and 830.11 minutes for $\rho = 0.80$). Because of the impact that pooling of

Table 2: Input Parameters for Experimental Design Parameters and Threshold Policies for each Scenario.

Scenarios						Policies						
	V	ρ	$\overline{\lambda_r}$	$\overline{\lambda_w}$	e_r/e_w	1. Ave.	2. SIPP	3. PSA	4. ASA	5. Max	6. Min	7. None
1	25	0.4	2.50	2.50	100	2	(2, 2, 3, 3)	2	2	3	2	0
2	25	0.6	3.75	3.75	100	3	(2, 3, 3, 3)	3	3	3	2	0
3	25	0.8	5.00	5.00	100	3	(2, 3, 1, 0)	2	1	3	0	0
4	75	0.4	7.50	7.50	100	2	(2, 2, 3, 3)	2	2	3	2	0
5	75	0.6	11.25	11.25	100	3	(2, 3, 4, 4)	3	3	4	2	0
6	75	0.8	15.00	15.00	100	4	(3, 4, 3, 0)	3	2	4	0	0
7	100	0.4	10.00	10.00	100	2	(2, 2, 3, 3)	2	2	3	2	0
8	100	0.6	15.00	15.00	100	3	(2, 3, 4, 2)	3	3	4	2	0
9	100	0.8	20.00	20.00	100	4	(3, 4, 4, 0)	3	3	4	0	0
10	25	0.4	4.00	1.00	100	2	(2, 2, 3, 3)	2	2	3	2	0
11	25	0.6	6.00	1.50	100	3	(2, 3, 4, 1)	3	2	4	1	0
12	25	0.8	8.00	2.00	100	3	(3, 4, 0, 0)	2	1	4	0	0
13	75	0.4	12.00	3.00	100	2	(2, 2, 3, 4)	3	3	4	2	0
14	75	0.6	18.00	4.50	100	4	(2, 3, 5, 2)	3	3	5	2	0
15	75	0.8	24.00	6.00	100	5	(3, 5, 0, 0)	2	1	5	0	0
16	100	0.4	16.00	4.00	100	2	(2, 2, 3, 4)	3	3	4	2	0
17	100	0.6	24.00	6.00	100	4	(2, 3, 5, 3)	3	3	5	2	0
18	100	0.8	32.00	8.00	100	6	(3, 5, 0, 0)	2	1	5	0	0
19	25	0.4	2.50	2.50	1000	4	(3, 3, 4, 5)	4	3	5	3	0
20	25	0.6	3.75	3.75	1000	5	(3, 4, 5, 4)	4	4	5	3	0
21	25	0.8	5.00	5.00	1000	4	(4, 4, 2, 0)	3	2	4	0	0
22	75	0.4	7.50	7.50	1000	4	(3, 4, 4, 6)	4	4	6	3	0
23	75	0.6	11.25	11.25	1000	5	(3, 5, 6, 8)	5	5	8	3	0
24	75	0.8	15.00	15.00	1000	6	(4, 6, 5, 0)	4	3	6	0	0
25	100	0.4	10.00	10.00	1000	4	(3, 4, 4, 6)	4	4	6	3	0
26	100	0.6	15.00	15.00	1000	5	(3, 5, 6, 8)	5	5	8	3	0
27	100	0.8	20.00	20.00	1000	6	(4, 6, 6, 0)	5	4	6	0	0
28	25	0.4	4.00	1.00	1000	4	(3, 4, 5, 7)	4	4	7	3	0
29	25	0.6	6.00	1.50	1000	6	(4, 6, 6, 2)	5	4	6	2	0
30	25	0.8	8.00	2.00	1000	6	(4, 6, 0, 0)	3	2	6	0	0
31	75	0.4	12.00	3.00	1000	4	(3, 4, 6, 8)	5	5	8	3	0
32	75	0.6	18.00	4.50	1000	7	(4, 6, 10, 5)	6	6	10	4	0
33	75	0.8	24.00	6.00	1000	10	(4, 9, 0, 0)	4	2	9	0	0
34	100	0.4	16.00	4.00	1000	4	(3, 4, 6, 8)	5	5	8	3	0
35	100	0.6	24.00	6.00	1000	7	(4, 6, 10, 6)	7	6	10	4	0
36	100	0.8	32.00	8.00	1000	10	(4, 9, 0, 0)	4	2	9	0	0

Table 3: Performance Metrics from the Simulation Study Evaluating Threshold Policies for Non-Stationary Demand (Scenarios 1 - 9). (Note, the values in bold denote the policy that achieves the minimum weighted waiting cost for each scenario).

	V	ρ	$\overline{\lambda_r}$	$\overline{\lambda_w}$	e_r/e_w	Policy	k	$\mathbb{E}[W_r^k]$	$\mathbb{E}[W_w^k]$	θ_k
1	25	0.4	2.50	2.50	100	1, 3, 4, 6	2	0.00	0.23	**0.35**
1	25	0.4	2.50	2.50	100	2	(2, 2, 3, 3)	0.00	0.48	0.48
1	25	0.4	2.50	2.50	100	5	3	0.00	0.52	0.52
1	25	0.4	2.50	2.50	100	7	0	0.02	0.03	1.58
2	25	0.6	3.75	3.75	100	1, 3, 4, 5	3	0.86	52.79	**138.80**
2	25	0.6	3.75	3.75	100	2	(2, 3, 3, 3)	1.20	44.64	165.12
2	25	0.6	3.75	3.75	100	6	2	1.82	33.45	215.65
2	25	0.6	3.75	3.75	100	7	0	4.34	8.59	442.52
3	25	0.8	5.00	5.00	100	1, 5	3	19.03	858.40	**2761.10**
3	25	0.8	5.00	5.00	100	3	2	28.77	552.45	3429.52
3	25	0.8	5.00	5.00	100	4	1	43.45	386.16	4730.71
3	25	0.8	5.00	5.00	100	2	(2, 3, 1, 0)	47.98	524.90	5323.05
3	25	0.8	5.00	5.00	100	6, 7	0	62.98	260.96	6558.75
4	75	0.4	7.50	7.50	100	1, 3, 4, 6	2	0.00	0.00	0.00
4	75	0.4	7.50	7.50	100	2	(2, 2, 3, 3)	0.00	0.00	0.00
4	75	0.4	7.50	7.50	100	5	3	0.00	0.00	0.00
4	75	0.4	7.50	7.50	100	7	0	0.00	0.00	0.00
5	75	0.6	11.25	11.25	100	2	(2, 3, 4, 4)	0.00	0.31	**0.38**
5	75	0.6	11.25	11.25	100	1, 3, 4	3	0.00	0.19	0.50
5	75	0.6	11.25	11.25	100	6	2	0.00	0.12	0.58
5	75	0.6	11.25	11.25	100	5	4	0.01	0.45	1.21
5	75	0.6	11.25	11.25	100	7	0	0.03	0.05	2.64
6	75	0.8	15.00	15.00	100	1, 5	4	1.96	115.70	**312.03**
6	75	0.8	15.00	15.00	100	3	3	2.81	91.10	372.46
6	75	0.8	15.00	15.00	100	4	2	3.95	71.39	466.18
6	75	0.8	15.00	15.00	100	2	(3, 4, 3, 0)	5.30	77.74	607.45
6	75	0.8	15.00	15.00	100	6, 7	0	9.03	42.18	945.27
7	100	0.4	10.00	10.00	100	1, 3, 4, 6	2	0.00	0.00	0.00
7	100	0.4	10.00	10.00	100	2	(2, 2, 3, 3)	0.00	0.00	0.00
7	100	0.4	10.00	10.00	100	5	3	0.00	0.00	0.00
7	100	0.4	10.00	10.00	100	7	0	0.00	0.00	0.00
8	100	0.6	15.00	15.00	100	7	0	0.00	0.00	**0.00**
8	100	0.6	15.00	15.00	100	2	(2, 3, 4, 2)	0.00	0.01	0.01
8	100	0.6	15.00	15.00	100	1, 3, 4	3	0.00	0.01	0.01
8	100	0.6	15.00	15.00	100	5	4	0.00	0.01	0.01
8	100	0.6	15.00	15.00	100	6	2	0.00	0.00	0.01
9	100	0.8	20.00	20.00	100	1, 5	4	1.02	58.61	**160.40**
9	100	0.8	20.00	20.00	100	3, 4	3	1.39	45.91	185.40
9	100	0.8	20.00	20.00	100	2	(3, 4, 4, 0)	2.90	49.65	339.67
9	100	0.8	20.00	20.00	100	6, 7	0	5.11	25.82	536.82

Table 4: Performance Metrics from the Simulation Study Evaluating Threshold Policies for Non-Stationary Demand (Scenarios 10-18). (Note, the values in bold denote the policy that achieves the minimum weighted waiting cost for each scenario).

	V	ρ	$\overline{\lambda_r}$	$\overline{\lambda_w}$	e_r/e_w	Policy	k	$\mathbb{E}[W_r^k]$	$\mathbb{E}[W_w^k]$	θ_k
10	25	0.4	4.00	1.00	100	1, 3, 4, 6	2	0.07	0.70	**7.99**
10	25	0.4	4.00	1.00	100	7	0	0.10	0.19	9.99
10	25	0.4	4.00	1.00	100	2	(2, 2, 3, 3)	0.09	1.37	10.11
10	25	0.4	4.00	1.00	100	5	3	0.09	1.52	10.23
11	25	0.6	6.00	1.50	100	5	4	6.34	137.06	**771.26**
11	25	0.6	6.00	1.50	100	4	2	8.39	58.24	897.11
11	25	0.6	6.00	1.50	100	1, 3	3	8.49	89.51	938.49
11	25	0.6	6.00	1.50	100	6	1	9.05	33.68	938.50
11	25	0.6	6.00	1.50	100	2	(2, 3, 4, 1)	8.72	77.63	949.15
11	25	0.6	6.00	1.50	100	7	0	10.94	23.00	1116.62
12	25	0.8	8.00	2.00	100	1	3	119.01	1220.38	**13121.63**
12	25	0.8	8.00	2.00	100	3	2	124.73	865.99	13339.17
12	25	0.8	8.00	2.00	100	5	4	119.26	1719.06	13645.31
12	25	0.8	8.00	2.00	100	4	1	143.21	768.02	15089.13
12	25	0.8	8.00	2.00	100	2	(3, 4, 0, 0)	142.96	842.52	15138.80
12	25	0.8	8.00	2.00	100	6, 7	0	151.83	547.39	15730.41
13	75	0.4	12.00	3.00	100	1, 6	2	0.00	0.00	0.00
13	75	0.4	12.00	3.00	100	2	(2, 2, 3, 4)	0.00	0.00	0.00
13	75	0.4	12.00	3.00	100	3, 4	3	0.00	0.00	0.00
13	75	0.4	12.00	3.00	100	5	4	0.00	0.00	0.00
13	75	0.4	12.00	3.00	100	7	0	0.00	0.00	0.00
14	75	0.6	18.00	4.50	100	3, 4	3	0.12	1.73	**13.40**
14	75	0.6	18.00	4.50	100	5	5	0.10	3.22	13.62
14	75	0.6	18.00	4.50	100	1	4	0.13	2.35	15.36
14	75	0.6	18.00	4.50	100	2	(2, 3, 5, 2)	0.14	2.33	16.42
14	75	0.6	18.00	4.50	100	6	2	0.17	1.15	18.35
14	75	0.6	18.00	4.50	100	7	0	0.23	0.71	23.45
15	75	0.8	24.00	6.00	100	1, 5	5	30.80	356.92	**3436.50**
15	75	0.8	24.00	6.00	100	3	2	36.39	237.64	3876.80
15	75	0.8	24.00	6.00	100	4	1	38.56	215.62	4071.35
15	75	0.8	24.00	6.00	100	2	(3, 5, 0, 0)	41.41	226.07	4366.96
15	75	0.8	24.00	6.00	100	6, 7	0	43.22	183.22	4505.13
16	100	0.4	16.00	4.00	100	1, 6	2	0.00	0.00	0.00
16	100	0.4	16.00	4.00	100	2	(2, 2, 3, 4)	0.00	0.00	0.00
16	100	0.4	16.00	4.00	100	3, 4	3	0.00	0.00	0.00
16	100	0.4	16.00	4.00	100	5	4	0.00	0.00	0.00
16	100	0.4	16.00	4.00	100	7	0	0.00	0.00	0.00
17	100	0.6	24.00	6.00	100	3, 4	3	0.01	0.21	**1.19**
17	100	0.6	24.00	6.00	100	1	4	0.01	0.25	1.21
17	100	0.6	24.00	6.00	100	6	2	0.01	0.08	1.24
17	100	0.6	24.00	6.00	100	2	(2, 3, 5, 3)	0.01	0.41	1.50
17	100	0.6	24.00	6.00	100	7	0	0.02	0.03	1.75
17	100	0.6	24.00	6.00	100	5	5	0.01	0.54	1.95
18	100	0.8	32.00	8.00	100	1	6	20.19	282.95	**2301.81**
18	100	0.8	32.00	8.00	100	5	5	21.13	260.21	2373.33
18	100	0.8	32.00	8.00	100	3	2	25.87	182.58	2769.40
18	100	0.8	32.00	8.00	100	4	1	28.10	159.89	2969.84
18	100	0.8	32.00	8.00	100	2	(3, 5, 0, 0)	29.75	167.23	3142.18
18	100	0.8	32.00	8.00	100	6, 7	0	30.29	139.58	3168.45

Table 5: Performance Metrics from the Simulation Study Evaluating Threshold Policies for Non-Stationary Demand (Scenarios 19-27). (Note, the values in bold denote the policy that achieves the minimum weighted waiting cost for each scenario).

	V	ρ	$\overline{\lambda_r}$	$\overline{\lambda_w}$	e_r/e_w	Policy	k	$\mathbb{E}[W_r^k]$	$\mathbb{E}[W_w^k]$	θ_k
19	25	0.4	2.50	2.50	1000	4, 6	3	0.00	0.52	**0.52**
19	25	0.4	2.50	2.50	1000	1, 3	4	0.00	1.24	1.24
19	25	0.4	2.50	2.50	1000	2	(3, 3, 4, 5)	0.00	1.28	1.99
19	25	0.4	2.50	2.50	1000	5	5	0.00	2.86	2.86
19	25	0.4	2.50	2.50	1000	7	0	0.02	0.03	15.54
20	25	0.6	3.75	3.75	1000	1, 5	5	0.34	145.19	**486.32**
20	25	0.6	3.75	3.75	1000	2	(3, 4, 5, 4)	0.68	94.93	774.32
20	25	0.6	3.75	3.75	1000	3, 4	4	0.70	80.77	776.55
20	25	0.6	3.75	3.75	1000	6	3	0.86	52.79	912.83
20	25	0.6	3.75	3.75	1000	7	0	4.34	8.59	4347.91
21	25	0.8	5.00	5.00	1000	1, 5	4	13.64	1652.54	**15296.28**
21	25	0.8	5.00	5.00	1000	3	3	19.03	858.40	19885.43
21	25	0.8	5.00	5.00	1000	4	2	28.77	552.45	29323.14
21	25	0.8	5.00	5.00	1000	2	(4, 4, 2, 0)	49.16	1041.35	50198.87
21	25	0.8	5.00	5.00	1000	6, 7	0	62.98	260.96	63238.86
22	75	0.4	7.50	7.50	1000	1, 3, 4	4	0.00	0.00	0.00
22	75	0.4	7.50	7.50	1000	2	(3, 4, 4, 6)	0.00	0.00	0.00
22	75	0.4	7.50	7.50	1000	5	6	0.00	0.00	0.00
22	75	0.4	7.50	7.50	1000	6	3	0.00	0.00	0.00
22	75	0.4	7.50	7.50	1000	7	0	0.00	0.00	0.00
23	75	0.6	11.25	11.25	1000	1, 3, 4	5	0.00	0.43	**0.75**
23	75	0.6	11.25	11.25	1000	5	8	0.00	1.61	1.61
23	75	0.6	11.25	11.25	1000	2	(3, 5, 6, 8)	0.00	1.03	1.64
23	75	0.6	11.25	11.25	1000	6	3	0.00	0.19	3.28
23	75	0.6	11.25	11.25	1000	7	0	0.03	0.05	25.90
24	75	0.8	15.00	15.00	1000	1, 5	6	0.98	183.48	**1158.64**
24	75	0.8	15.00	15.00	1000	3	4	1.96	115.70	2079.09
24	75	0.8	15.00	15.00	1000	4	3	2.81	91.10	2904.67
24	75	0.8	15.00	15.00	1000	2	(4, 6, 5, 0)	5.11	114.27	5225.39
24	75	0.8	15.00	15.00	1000	6, 7	0	9.03	42.18	9073.01
25	100	0.4	10.00	10.00	1000	1, 3, 4	4	0.00	0.00	0.00
25	100	0.4	10.00	10.00	1000	2	(3, 4, 4, 6)	0.00	0.00	0.00
25	100	0.4	10.00	10.00	1000	5	6	0.00	0.00	0.00
25	100	0.4	10.00	10.00	1000	6	3	0.00	0.00	0.00
25	100	0.4	10.00	10.00	1000	7	0	0.00	0.00	0.00
26	100	0.6	15.00	15.00	1000	7	0	0.00	0.00	**0.01**
26	100	0.6	15.00	15.00	1000	6	3	0.00	0.01	0.01
26	100	0.6	15.00	15.00	1000	1, 3, 4	5	0.00	0.02	0.02
26	100	0.6	15.00	15.00	1000	2	(3, 5, 6, 8)	0.00	0.07	0.07
26	100	0.6	15.00	15.00	1000	5	8	0.00	0.22	0.22
27	100	0.8	20.00	20.00	1000	1, 5	6	0.48	83.29	**567.56**
27	100	0.8	20.00	20.00	1000	3	5	0.75	74.28	821.07
27	100	0.8	20.00	20.00	1000	4	4	1.02	58.61	1076.54
27	100	0.8	20.00	20.00	1000	2	(4, 6, 6, 0)	2.68	60.84	2736.44
27	100	0.8	20.00	20.00	1000	6, 7	0	5.11	25.82	5135.83

Table 6: Performance Metrics from the Simulation Study Evaluating Threshold Policies for Non-Stationary Demand (Scenarios 28-36). (Note, the values in bold denote the policy that achieves the minimum weighted waiting cost for each scenario).

	V	ρ	$\overline{\lambda_r}$	$\overline{\lambda_w}$	e_r/e_w	Policy	k	$\mathbb{E}[W_r^k]$	$\mathbb{E}[W_w^k]$	θ_k
28	25	0.4	4.00	1.00	1000	5	7	0.02	16.58	**35.00**
28	25	0.4	4.00	1.00	1000	1, 3, 4	4	0.04	2.40	38.20
28	25	0.4	4.00	1.00	1000	2	(3, 4, 5, 7)	0.06	7.41	70.81
28	25	0.4	4.00	1.00	1000	6	3	0.09	1.52	88.62
28	25	0.4	4.00	1.00	1000	7	0	0.10	0.19	98.13
29	25	0.6	6.00	1.50	1000	1, 5	6	5.09	320.55	**5410.27**
29	25	0.6	6.00	1.50	1000	4	4	6.34	137.06	6479.08
29	25	0.6	6.00	1.50	1000	3	5	7.21	209.68	7423.47
29	25	0.6	6.00	1.50	1000	6	2	8.39	58.24	8447.01
29	25	0.6	6.00	1.50	1000	2	(4, 6, 6, 2)	8.85	196.40	9049.32
29	25	0.6	6.00	1.50	1000	7	0	10.94	23.00	10959.16
30	25	0.8	8.00	2.00	1000	1, 5	6	102.31	4093.18	**106403.38**
30	25	0.8	8.00	2.00	1000	3	3	119.01	1220.38	120232.89
30	25	0.8	8.00	2.00	1000	4	2	124.73	865.99	125597.76
30	25	0.8	8.00	2.00	1000	6, 7	0	151.83	547.39	152377.63
30	25	0.8	8.00	2.00	1000	2	(4, 6, 0, 0)	153.51	1109.95	154619.47
31	75	0.4	12.00	3.00	1000	1	4	0.00	0.00	0.00
31	75	0.4	12.00	3.00	1000	2	(3, 4, 6, 8)	0.00	0.00	0.00
31	75	0.4	12.00	3.00	1000	3, 4	5	0.00	0.00	0.00
31	75	0.4	12.00	3.00	1000	5	8	0.00	0.00	0.00
31	75	0.4	12.00	3.00	1000	6	3	0.00	0.00	0.00
31	75	0.4	12.00	3.00	1000	7	0	0.00	0.00	0.00
32	75	0.6	18.00	4.50	1000	5	10	0.06	15.83	**74.41**
32	75	0.6	18.00	4.50	1000	3, 4	6	0.08	5.06	82.25
32	75	0.6	18.00	4.50	1000	2	(4, 6, 10, 5)	0.09	8.94	94.42
32	75	0.6	18.00	4.50	1000	1	7	0.12	7.32	131.58
32	75	0.6	18.00	4.50	1000	6	4	0.13	2.35	132.48
32	75	0.6	18.00	4.50	1000	7	0	0.23	0.71	228.16
33	75	0.8	24.00	6.00	1000	5	9	24.74	608.27	**25350.98**
33	75	0.8	24.00	6.00	1000	1	10	25.02	737.99	25759.68
33	75	0.8	24.00	6.00	1000	3	4	32.89	322.93	33208.57
33	75	0.8	24.00	6.00	1000	4	2	36.39	237.64	36629.23
33	75	0.8	24.00	6.00	1000	2	(4, 9, 0, 0)	42.11	264.77	42377.40
33	75	0.8	24.00	6.00	1000	6, 7	0	43.22	183.22	43402.40
34	100	0.4	16.00	4.00	1000	1	4	0.00	0.00	0.00
34	100	0.4	16.00	4.00	1000	2	(3, 4, 6, 8)	0.00	0.00	0.00
34	100	0.4	16.00	4.00	1000	3, 4	5	0.00	0.00	0.00
34	100	0.4	16.00	4.00	1000	5	8	0.00	0.00	0.00
34	100	0.4	16.00	4.00	1000	6	3	0.00	0.00	0.00
34	100	0.4	16.00	4.00	1000	7	0	0.00	0.00	0.00
35	100	0.6	24.00	6.00	1000	1, 3	7	0.00	1.01	**5.99**
35	100	0.6	24.00	6.00	1000	6	4	0.01	0.25	9.84
35	100	0.6	24.00	6.00	1000	7	0	0.02	0.03	17.29
35	100	0.6	24.00	6.00	1000	4	6	0.02	0.96	19.34
35	100	0.6	24.00	6.00	1000	5	10	0.02	2.54	19.36
35	100	0.6	24.00	6.00	1000	2	(4, 6, 10, 6)	0.02	2.20	21.81
36	100	0.8	32.00	8.00	1000	1	10	16.77	447.59	**17219.31**
36	100	0.8	32.00	8.00	1000	5	9	17.84	413.06	18251.78
36	100	0.8	32.00	8.00	1000	3	4	22.38	229.84	22607.44
36	100	0.8	32.00	8.00	1000	4	2	25.87	182.58	26050.75
36	100	0.8	32.00	8.00	1000	6, 7	0	30.29	139.58	30428.25
36	100	0.8	32.00	8.00	1000	2	(4, 9, 0, 0)	30.83	191.99	31024.85

vehicles has on reducing the impact of variability, the average waiting time also varies with number of vehicles available (average reserve customer waiting times are 22.23, 4.89, and 2.96 minutes for $V = 25$, 75, and 100, respectively; average walk-in customer waiting times are 708.18, 106.89, and 67.20 minutes for $V = 25$, 75, and 100, respectively). Scenarios with large walk-in customer waiting times can occur for two primary reasons. First, the use of a threshold policy means that both walk-in customers and vehicles can wait simultaneously. Second, the non-stationary arrival patterns create peak periods when arrival rates are high and waiting times are large, and non-peak periods when vehicles are waiting idle due to low customer arrival rates.

Our numerical testing provides insights into recommending threshold policies in a dynamic non-stationary environment. A threshold policy outperforms the no threshold policy in all but two cases (i.e., Scenarios 8 and 26 have policy 7 performing best; however, both scenarios represent systems that have close to zero expected waiting times for both customer classes). Therefore, implementing a threshold policy can improve profitability of rental providers. We also find that simple, static threshold policies tend to outperform more complicated, dynamic threshold policies, even when reserve customer arrival demand is non-stationary. Out of all of the policies tested, the SIPP approach (Policy 2) is the only dynamic threshold policy in that K^* is adjusted depending on the customer arrival rate at a particular time. From our analysis, we find that a simple static policy outperforms a more complicated, dynamic policy (i.e., Policy 2 is the recommended policy only for scenario 5). The SIPP approach does not perform well when the system utilization is high, which can be attributed to violating many of the assumptions that are made with applying independent queuing models to set threshold quantities. These violated assumptions include that a system achieves steady state and that waiting times in adjacent periods are statistically independent [11].

A policy that uses the average arrival rate (Policy 1) tends to perform well (i.e., Policy 1 performed best in 26/36 scenarios). However, not one single policy dominates in all scenarios. Therefore, consideration should be given to the number of vehicles, fleet utilization, arrival rates, and penalty values. When reserve customer waiting time has a much greater penalty than walk-in customer waiting times (i.e., when $e_r/e_w = 1000$), a policy that uses the maximum threshold or average arrival rate policy (Policy 1 or 5) is recommended.

6. Conclusions

Because both responsiveness and customer differentiation impact profitability, we developed stochastic models to analyze rental provider policies that have two classes of customers. We analyzed a condition-based threshold policy that provides vehicles to walk-in customers only if the number of vehicles available exceeds a threshold quantity that considered the waiting times of both customer classes. We modeled such a rental depot as a multi-class non-work-conserving semi-open queuing with stochastic inputs. We developed an optimization model to determine the optimal threshold quantities that considers customer waiting times and relative importance for both customer classes. When arrival rates are stationary, we analyzed threshold policies using closed-form expressions for both exponential and deterministic vehicle unavailability periods. For non-stationary demand rates, we developed seven policies and analyzed them using a discrete-event simulation model. To provide insights into threshold policies, we conducted numerical testing that varied the number of vehicles, fleet utilization, arrival rates, and penalty weights.

Through the development of our analytical models and our numerical results we were able to provide the following non-intuitive managerial insights about rental provider threshold policies.

- Although a threshold policy is implemented to reduce the probability that a reserve customer may have to wait because all the vehicles are rented, the optimal threshold quantity does not always increase as the reserve customer arrival rate increases. Instead, because of the nonlinear relationship between utilization and customer waiting time, the optimal threshold quantity is a concave function in reserve customer arrival rates. Similarly, the optimal threshold quantity does not always increase as the walk-in arrival rate increases.
- We identify that the threshold quantities are fairly insensitive to the distribution of vehicle unavailability times.
- Through numerical testing we find that simple, static threshold policies tend to outperform more complicated, dynamic threshold policies, even when reserve customer arrival demand is non-stationary.

Our models can be adapted in other situations where customer classes with different priorities compete for resource capacity, such as berth assignment at a container terminal or shelf-space assignment at a third-party logistics provider. Future research could also examine non-stationary demand threshold policies where multiple classes of vehicles exist. Such an analysis could include the trade-off between the cost of upgrading a customer to a higher vehicle class and the decrease in waiting time costs.

Acknowledgments

The authors would like to thank representatives from the vehicle rental provider who participated in discussions of their operations and provided inputs for this research.

References

[1] Altman, E., Jimenez, T., and Koole, G., "On Optimal Call Admission Control in a Resource-Sharing System," *IEEE Transactions on Communications*, 49, 9, 1659–1668 (2001).

[2] Auto Rental News, "2011 U.S. Car Rental Market, Auto Rental News, Fact Book 2012," (2011), $http://www.autorentalnews.com/fc_resources/Editorial/ARN-6.pdf$.

[3] Barcelo, F., Casares, V., and Paradells, J., "M/D/C Queue with Priority: Application to Trunked Mobile Radio Systems," *Electronics Letters*, 32, 1644–1645 (1996).

[4] Brown, C., "Turning a Profit," *Rental Operations* (2011), http://www.autorentalnews.com/Article/Story/2011/04/Turning-A-Profit/Page/1.aspx.

[5] Busse, M., and Swinkels, J., "Enterprise Rent-A-Car," Technical Report KEL612, Kellogg School of Management (2012).

[6] Carroll, W. J., and Grimes, R. C., "Evolutionary Change in Product Mangement: Experiences in the Car Rental Industry," *Interfaces*, 25, 5, 84–104 (1995).

[7] Fink, A., and Reiners, T., "Modeling and Solving the Short-Term Car Rental Logistics Problem," *Transportation Research Part E*, 42, 272–292 (2006).

[8] Gans, N., and Savin, S., "Pricing and Capacity Rationing for Rentals with Uncertain Durations," *Management Science*, 53, 3, 390–407 (2007).

[9] George, D., and Xia, C. H., "Fleet-Sizing and Service Availability for a Vehicle Rental System via Closed Queueing Networks," *European Journal of Operational Research*, 211, 1, 198–207 (2011).

[10] Green, L., and Kolesar, P., "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals," *Management Science*, 37, 1, 84–97 (1991).

[11] Green, L., Kolesar, P., and Soares, J., "Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demands," *Operations Research*, 49, 4, 549–564 (2001).

[12] Guide, V., Souza, G., and Van Der Laan, E., "Performance of Static Priority Rules for Shared Facilities in a Remanufacturing Shop with Disassembly and Reassembly," *European Journal of Operational Research*, 164, 2, 341–353 (2005).

[13] Helm, J. E., AhmadBeygi, S., and Van Oyen, M. P., "Design and Analysis of Hospital Admission Control for Operational Effectiveness," *Production and Operations Management*, 20, 3 (2011).

[14] HighBeam Business, "Passenger Car Rental Industry Report," Technical Report SIC 7514, HighBeam Business (2012), http://business.highbeam.com/industry-reports/personal/passenger-car-rental.

[15] Koutras, V., Platis, A., and Gravvanis, G., "Optimal Server Resource Reservation Policies for Priority Classes of Users Under Cyclic Non-Homogeneous Markov Modeling," *European Journal of Operational Research*, 198, 2, 545–556 (2009).

[16] Lee, D.-S., and Sengupta, B., "Queueing Analysis of a Threshold Based Priority Scheme for ATM Networks," *IEEE/ACM Transactions on Networking*, 1, 6, 709–717 (1993).

[17] Miller, B. L., "A Queuing Reward System with Several Customer Classes," *Management Science*, 16, 3, 234–245 (1969).

[18] Ormeci, L., Burnetas, A., and van der Wal, J., "Admissions Policies to a Two-Class Loss System," *Stochastic Models*, 17, 513–539 (2001).

[19] Papier, F., and Thonemann, U. W., "Queuing Models for Sizing and Structuring Rental Fleets," *Transportation Science*, 42, 3, 302–317 (2008).

[20] Papier, F., and Thonemann, U. W., "Capacity Rationing in Stochastic Rental Systems with Advance Demand Information," *Operations Research*, 58, 274–288 (2010).

[21] Papier, F., and Thonemann, U. W., "Capacity Rationing in Rental Systems with Two Customer Classes and Batch Arrivals," *Omega*, 39, 73–85 (2011).

[22] Priceline.com, “New Priceline.com App Data Suggests That Rental Car Customers May Be The Most Last-Minute Bookers Of Them All,” (2012), http://www.prnewswire.com/news-releases-test/new-pricelinecom-app-data-suggests-that-rental-car-customers-may-be-the-most-last-minute-bookers-of-them-all-148569005.html.

[23] Rice, K., “J.D. Powers Says Consumers Are Happier With Car Rentals, http://www.travelpulse.com/jd-powers-says-consumers-are-happier-with-car-rentals.html,” (2011).

[24] Savin, S. V., Cohen, M. A., Gans, N., and Katalan, Z., “Capacity Management in Rental Businesses with Two Customer Bases,” *Operations Research*, 53, 617–631 (2005).

[25] Schaack, C., and Larson, R. C., “An N-Server Cutoff Priority Queue,” *Operations Research*, 34, 2, 257–266 (1986).

[26] Shimshak, D., Gropp Damico, D., and Burden, H., “A Priority Queuing Model of a Hospital Pharmacy Unit,” *European Journal of Operational Research*, 7, 4, 350–354 (1981).

[27] Shonick, W., and Jackson, J. R., “An Improved Stochastic Model for Occupancy-Related Random Variables in General Acute Hospitals,” *Operations Research*, 21, 4, 952–965 (1973).

[28] Sleptchenko, A., Van Der Heijden, M., and Van Harten, A., “Using Repair Priorities to Reduce Stock Investment in Spare Part Networks,” *European Journal of Operational Research*, 163, 3, 733–750 (2005).

[29] Taylor, I. D. S., and Templeton, J. G. C., “Waiting Time in a Multi-Server Cutoff-Priority Queue, and Its Application to an Urban Ambulance Service,” *Operations Research*, 28, 5, 1168–1188 (1980).

[30] Yang, Y., Jin, W., and Hao, X., “Car Rental Logistics Problem: A Review of Literature,” *IEEE International Conference on Service Operations and Logistics, and Informatics*, 2, 2815–2819 (2008).