INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Emergency Medical Service System Design under Service Level Constraints for Heterogeneous Patients

Sachin Jayaswal

W.P. No. 2014-11-04

November 2014

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

Emergency Medical Service System Design under Service Level Constraints for Heterogeneous Patients

Sachin Jayaswal

Production & Quantitative Methods
Indian Institute of Management Ahmedabad
srikumark@iimahd.ernet.in

Abstract

We study the problem of locating Emergency Medical Service (EMS) facilities in the presence of service level constraints for patients with acuity levels ranging from resuscitation to non-urgent. Each patient arriving at any EMS facility is triaged as either resuscitation/high priority or less urgent/low priority, where high priority patients are always served on a priority basis. The problem is to optimally locate EMS facilities and allocate their service zones to satisfy the following coverage and service level constraints: (i) each user zone is served by an EMS facility that is within a given coverage radius; (ii) at least α^h proportion of the resuscitation cases at any EMS facility should be admitted immediately without having to wait; (iii) at least α^l proportion of the cases belonging to low priority class at any EMS facility should not have to wait for more than τ^l minutes. For this, we model the network of EMS facilities as spatially distributed M/M/1 priority queues, whose locations and user allocations need to be determined. The resulting integer programming problem is challenging to solve, especially in absence of any known analytical expression for the waiting time distribution of low priority customers in an M/M/1 priority queue. We develop a cutting plane based solution algorithm, exploiting the concavity of the waiting time distribution of low priority customers to approximate its non-linearity using tangent planes, determined numerically using matrix geometric method. Using a case study of locating EMS facilities in Austin, Texas, we present computational results and managerial insights.

Keywords: Location; Congestion; Service Level Constraints; Cutting Plane; Priority Queue

Emergency Medical Service System Design under Service Level Constraints for Heterogeneous Patients

1 Introduction

Emergency Care Facilities, Urgent Care Centers or Emergency Departments in Hospitals, henceforth collectively referred to as Emergency Medical Service (EMS) Facilities, are 24 hour, 7 days-a-week, medical facilities focussed on the delivery of ambulatory care to treat injuries or illnesses requiring immediate care. They are equipped with all the appropriate infrastructure (human resources, equipments and technologies) required for the assessment, resuscitation, stabilization, and, where appropriate, either admission or transfer, of patients in need of urgent medical care.[1] However, overcrowding, and consequently long wait, is a common problem at EMS facilities. For example, U.S. emergency departments witnessed 123.8 million visits in 2008, of which only 18% were seen within 15 minutes, leaving the majority of them waiting (Gilboy et al., 2011). Many jurisdictions around the world have, therefore, developed acuity rating systems to help EMS service providers correctly triage patients, i.e., prioritize patients based on the acuity of their injury/illness, to ensure those who have the most urgent need get the first access to urgent care. Emergency departments in the U.S., for example, use a 5-level Emergency Severity Index (ESI) to triage their patients (Gilboy et al., 2011). Along similar lines, Canadian Triage and Acuity Scale (CTAS) guidelines also classify an emergency patient into one of five acuity levels, as shown in Table 1, based on the acuity of her injury/illness (Murray, 2003). They further prescribe, for each acuity level, a time standard within which at least a given proportion of patients in that level (expressed as "Performance indicator threshold" in Table 1) should be seen by a physician after triage. Similarly, Australian Triage Scale (ATS)[2] also classifies patients in emergency departments into one of five acuity levels, although with somewhat different performance indicator thresholds.

In this paper, we study a service system design problem in the context of EMS, by taking into

[1] Erie St. Clair Local Health Integration Network: www.eriestclairlhin.on.ca

[2] https://www.acem.org.au/getattachment/d19d5ad3-e1f4-4e4f-bf83-7e09cae27d76/G24-Implementation-of-the-Australasian-Triage-Scal.aspx

Table 1: Canadian Emergency Department Triage and Acuity Scale (CTAS)

Level	Acuity	Examples of symptoms	Max wait before treatment	Performance indicator threshold
1	Resuscitation	Cardiac and/or pulmonary arrest; Major trauma (multiple system injury)	Immediate	98%
2	Emergent	Visceral pain; Gastrointestinal bleed; acid/alkali exposure to eyes	15 min	95%
3	Urgent	Moderate trauma (fractures, dislocations); mild/moderate asthma	30 min	90%
4	Less urgent	Minor trauma (sprains, contusions); ear ache	1 hour	85%
5	Non urgent	sore throat; vomiting with no signs of diarrhea and age > 2	2 hours	80%

account the heterogeneity of the arriving patients, as discussed above, which requires different performance targets for different acuity levels. Motivated by the development of acuity rating systems (like ESI in US, CTAS in Canada, ATS in Australia), which are meant to help EMS service providers correctly triage patients into different acuity classes, we study the problem in the presence of heterogeneous patients, belonging to different classes. To the best of our knowledge, this is the first study on a location-allocation problem in presence of heterogeneous customers with a different service level requirement for each class, and where the service level constraint for each customer class is defined using the complete distribution of its waiting time, as opposed to its average waiting time at an EMS facility. We model the network of EMS facilities as spatially distributed M/M/1 priority queues, whose locations and user allocations need to be determined. The resulting integer programming problem is challenging to solve, especially in absence of any known analytical expression for the waiting time distribution of low priority customers in an M/M/1 priority queue. We develop a cutting plane based solution algorithm, exploiting the concavity of the waiting time distribution of low priority customers to approximate its non-linearity using tangent planes, determined numerically using matrix geometric method. We present an illustrative case study using the 33-zone problem representing Austin, Texas at census tract level.

The remainder of the paper is organized as follows. We first present a review of the related literature, and identify our contribution in section 2. In section 3, we present a description of the problem, followed by its Integer Programming (IP) formulation. Section 4 presents the solution method. This is followed by illustrative examples, computational results and discussions, presented in section 5. Section 6 concludes the paper with a summary of results and directions for future research.

2 Related Literature

The problem considered in this paper belongs to the broad class of “Facility Location Problem with Stochastic Demand and Congestion” (Berman and Krass, 2002). The literature in this area can be categorised into two classes, depending on whether the service facility is mobile (e.g. ambulance) or immobile (e.g. EMS facilities, walk-in-clinics etc.). The EMS location-allocation problem that we study in this paper falls under the latter category of immobile servers. Boffey et al. (2007) provide a thorough review of the literature for the immobile server category. The literature in this category can be further divided depending on the way the issue of response time, in presence of congestion, is addressed. The first category of papers penalizes the service delay directly in the objective function. The model results in an IP with a non-linear objective function. Amiri (1997), Wang et al. (2002), Elhedhli (2006), Castillo et al. (2009), Vidyarthi and Jayaswal (2014), Vidyarthi and Kuzgunkaya (2014), among others, belong to this category.

The second category of literature on the problems with immobile servers imposes constraint(s) to ensure that waiting time, queue length or the proportion of demand lost due to congestion/insufficient coverage does not exceed a certain threshold. Marianov and Serra (1998) present a maximal covering location-allocation model to locate p centres and to allocate users to them such that maximum population is covered. The population at a user node is said to be covered if: (i) it is allocated to a center within a threshold time or distance; and (ii) a user from that node has to wait at its allocated centre with no more than b other people or no more than time τ with a probability of at least α. Marianov and Serra (2002) present a set covering version of this problem. Baron et al. (2008) study a similar problem under more relaxed assumptions, and with a constraint on the average waiting time at the service facilities. Berman et al. (2006) study a location-allocation problem wherein the demand may be lost either due to insufficient coverage, or due to the customers balking away on seeing a long queue at a facility. The objective of their study is to locate the minimum number of facilities required to ensure that the demand lost from either source does not exceed a given level. Zhang et al. (2012) study the impact of client choice (probabilistic versus closest facility) in a preventive healthcare facility location problem. They impose an upper bound on the mean waiting time, besides a lower bound on the workload requirement, at each open facility. Aboolian et al. (2012) study a location-allocation problem for preventive medical facilities, explicitly taking into account the sensitivity of demand to travel distance and congestion induced delays. The objective is to maximize profit (by private clinics providing preventive medical services) subject to the constraint that the expected waiting time of users should not exceed a given threshold.

All these papers cited above consider all service calls (arrivals) to be equally important in that they impose the same waiting time constraint, independent of their acuity levels. Such a constraint is inadequate/inappropriate in the context of EMS where calls may range from critical life-threatening to non-urgent stable. For example, CTAS, as discussed above, classifies the emergency department visits as one of the following in the decreasing order of acuity: (i) Resuscitation, (ii) Emergent, (ii) Urgent, (iv) Less urgent, and (v) Non-urgent. While a waiting time of one hour may be acceptable for Non-urgent cases, the same may prove fatal for resuscitation cases.

Silva and Serra (2008) present a more realistic model, which accounts for the priority of patients arriving at EMS facilities based on the acuity levels. For this, they model the network of EMS facilities, given their locations, as spatially distributed queuing facilities. Each queuing facility is assumed to serve its arriving patients using a priority discipline such that cases of higher acuity receive priority over those of lower acuity. They impose a different maximum average waiting time requirement for each priority class, and use a heuristic approach to deal with the complexity of the model. However, Triage and Acuity Scale guidelines seldom prescribe performance measures for EMS facilities based on average waiting time for any acuity level. They rather specify performance measures in terms of the proportion of patients in each acuity level that is served within a prescribed time standard (for example, refer to the Canadian Triage and Acuity Scale measures in Table 1).

In this paper, we present a location-allocation model for EMS systems with performance measure requirements that closely resemble what is prescribed in Triage and Acuity Rating guidelines. Accordingly, we present a probabilistic service level constraint for each patient class, defined using the complete distribution of its waiting time at an EMS facility. To the best of our knowledge, this is the first study on a location-allocation problem in presence of heterogeneous customers with a different service level requirement for each class, and where the service level constraint for each customer class is defined using the complete distribution of its waiting time, as opposed to its average waiting time at an EMS facility. We model the network of EMS facilities as spatially distributed priority M/M/1 queues, whose locations and user-allocations need to be decided. This makes the problem challenging to solve, especially in absence of any known analytical expression for waiting time distribution for low priority customers in a priority M/M/1 queue. We develop a cutting plane based solution algorithm, exploiting the concavity of the waiting time distribution of low priority customers to approximate its non-linearity using tangent planes, determined numerically using matrix geometric method.

3 Problem Description and Formulation

Consider a set of nodes I that represents the census tracts or census blocks, at which population is assumed to be aggregated. A subset $J \subseteq I$ of nodes also act as candidates sites for EMS facilities location. We assume that the number of patients at node $i \in I$ who require emergency medical services per unit time is random, described by a stationary Poisson process with mean λ_i. The assumption of Poisson demand arrivals does hold true for some data sets (Nair and Miller-Hooks, 2009). As discussed in section 1, these patients vary in their degree of acuity. However, for tractability, they are broadly classified in this paper only as resuscitation/high priority (denoted by h) that require immediate access to the emergency services, or less urgent/low priority (denoted by l), which subsumes all the remaining acuity levels.

At any EMS facility, patients from the same priority class are given access to physicians in the order of the arrival, i.e., First-Come-First-Served (FCFS). However, high priority (resuscitation) patients, because of the criticality of their cases, are always treated with priority compared to low priority patients. Priority given to resuscitation patients may be preemptive or non-preemptive. Under preemptive priority, if an EMS facility, at the arrival of a high priority patient, is completely busy, then a physician serving a low priority patient (e.g. a less urgent patient with a fractured leg) will put the current patient in wait and start attending the high priority patient. On the other hand, under non-preemptive priority, a high priority patient will still have to wait even though some of the physicians may be busy serving low priority patients. In this paper, we study both of these priority schemes. For the priority scheme to work, patients are assumed to be already triaged by paramedics in the ambulances that bring them to EMS facilities. Walk-in patients, on the other hand, are triaged, upon their arrival to EMS facilities, by triage nurses. We assume that there are enough triage nurses so that arriving patients can be immediately triaged.

Let f_i^h and f_i^l ($f_i^h + f_i^l = 1$) denote the proportions of service calls originating from node i that belong to high priority and low priority classes, respectively. Thus, the arrivals of high and low priority patients per unit time from node i can also be described as Poisson processes with means $\lambda_i^h = f_i^h \lambda_i$ and $\lambda_i^l = f_i^l \lambda_i$ respectively. Further, if $x_{ij}^c = 1$ indicates c^{th} patients of priority class c ($c \in \{h, l\}$) residing at node i patronizing facility j, then the arrival rate of patients at facility j can be expressed, using superposition of Poisson processes, as $\Lambda_j = \sum_{i \in I} \sum_{c \in \{h,l\}} \lambda_i^c x_{ij}^c$, while that for a given class c can be expressed as $\Lambda_j^c = \sum_{i \in I} \lambda_i^c x_{ij}^c$.

Acuity levels of different patients arriving at EMS facilities, unlike at preventive healthcare facilities, often differ, and so do their expected service times. Accordingly, we allow for patients of different acuity levels (priority classes) arriving at an EMS facility j to have different mean service rates. Let μ_j^c denote the mean service rate for patients of priority class c at location j. The service rate reflects the number of patients a facility can serve in a given time period. We assume the service times at each facility follow an exponential distribution. The assumption of exponential service time distribution is again corroborated, although with a fatter tail, by the data used by Nair and Miller-Hooks (2009). Thus, every EMS facility can be modelled as a priority M/M/1 queue with the mean service rate denoted by μ_j^c for class c patients. To state the problem mathematically, we first summarize the following notations.

Parameters:

λ_i^c : Mean demand rate from c^{th} priority class patients at node i (per hour)

μ_j^c : Mean service rate for c^{th} priority class patients at facility j (per hour)

d_{ij} : Travel time between demand node i and EMS facility location j (minutes)

R : Coverage radius such that user at node i is said to be covered by EMS facility j only if $d_{ij} \leq R$

τ^c : Threshold on the maximum waiting time for service at EMS facility for a patient of priority class c (minutes).

W_j^c : Actual waiting time of a patient of priority class c at EMS facility j (minutes)

α^c : Minimum required service level at an EMS facility for a patient of priority class c; $\mathrm{P}(W_j^c \leq \tau^c) \geq \alpha^c$

FC_j : Cost (amortized) of opening and operating an EMS facility at location j ($/hour)

TC : Cost per unit of travel time ($/minute)

Sets:

I : Set of demand zones, indexed by i, $i \in I$

J : Set of candidate sites for the location of EMS facilities, indexed by j, $j \in J$

C : Set of patient priority levels, indexed by c, $c \in C = \{h, l\}$

$J(i)$: $\{j | d_{ij} \leq R\}$

$I(j)$: $\{i | d_{ij} \leq R\}$

Variables:

y_j = 1 if an EMS facility is located at node j; 0 otherwise

x_{ij}^c = 1 if the demand zone i is allocated to EMS facility at j; 0 otherwise

Derived Variables:

Λ_j^c : Mean arrival rate of patients of class c at EMS facility j; $\Lambda_j^c = \sum_{i \in I} \lambda_i^c x_{ij}^c$

$S_j^c(\tau^c)$: Service level achieved at EMS facility j; $S_j^c(\tau^c) = P(W_j^c \leq \tau^c)$

The problem of the healthcare planner/decision maker is to determine the optimal location of EMS facilities among the nodes in J, and to allocate users at each node $i \in I$ to these facilities, such that all the user nodes are covered. A user node $i \in I$ is said to be covered when: (i) it is within the coverage radius R of an EMS facility, that is, $d_{ij} \leq R$, where d_{ij} is the travel time between the user node i and the EMS facility j; and (ii) waiting time W_j^c at an EMS facility j for patients of priority class c is within τ^c with a probability of at least α^c ($\alpha^c \in (0,1)$), i.e., $P(W_j^c \leq \tau^c) \geq \alpha^c$. The optimal location-allocation decision is defined with respect to the minimum total cost, which consists of the cost of opening and operating EMS facilities and the travel costs of patients from their respective locations to their allocated EMS facilities. This can be mathematically stated as:

$$[LAP]: \quad Min \quad \sum_{j \in J} FC_j y_j + TC \sum_{c \in C} \sum_{i \in I} \sum_{j \in J(i)} \lambda_i^c d_{ij} x_{ij}^c \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in J(i)} x_{ij}^c = 1 \qquad \forall i \in I; c \in C \tag{2}$$

$$x_{ij}^c \leq y_j \qquad \forall i \in I, j \in J(i); c \in C \tag{3}$$

$$\sum_{c \in C} \sum_{i \in I(j)} \frac{\lambda_i^c x_{ij}^c}{\mu_j^c} \leq y_j \qquad \forall j \in J \tag{4}$$

$$S_j^h\left(\tau^h = 0\right) = P(W_j^h \leq 0) \geq \alpha^h y_j \qquad \forall j \in J \tag{5}$$

$$S_j^l\left(\tau^l\right) = P(W_j^l \leq \tau^l) \geq \alpha^l y_j \qquad \forall j \in J \tag{6}$$

$$x_{ij}^c \in \{0,1\}, \quad y_j \in \{0,1\} \qquad \forall i \in I; j \in J(i); c \in C \tag{7}$$

The first term in the objective function (1) captures the (amortized per unit time) cost of opening and operating EMS facilities, while the second term captures the total travel costs of all patients in the network per unit time. Constraint set (2) ensures that all patients from a given priority class c at a given node i are allocated to only one EMS facility within the coverage radius R. Constraint set (3) states that patients from a priority class c at node i cannot be allocated to another node j unless there is an EMS facility positioned at j. Constraint set (4) is required for the stability of the queue at any open EMS facility. Constraint set (5) represents the service level requirement for high priority (resuscitation) patients at the open EMS facilities. Since resuscitation cases are critical, such cases, on arrival at an EMS facility, should be admitted immediately without having to wait (i.e., $\tau^h = 0$) with a probability of at least α^h. Constraint set (5), under preemptive and non-preemptive priority, can be expressed using (5-P) and (5-NP), respectively, as given below:

$$S_j^h\left(\tau^h=0\right)=1-\frac{\sum_{i\in I(j)}\lambda_i^h x_{ij}^h}{\mu_j^h}\geq \alpha^h y_j \qquad \forall j\in J \qquad \text{(5-P)}$$

$$S_j^h\left(\tau^h=0\right)=1-\sum_{c\in C}\frac{\sum_{i\in I(j)}\lambda_i^c x_{ij}^c}{\mu_j^c}\geq \alpha^h y_j \qquad \forall j\in J \qquad \text{(5-NP)}$$

The form of constraint set (5-P) arises from the fact that under preemptive priority, a high priority patient will not incur any wait at an EMS facility j if, upon arrival, it finds no other high priority patients ahead of it. The probability of this, by PASTA (Poisson Arrivals See Time Averages) property, is given by $1-\Lambda_j^h/\mu_j^h = 1-\sum_{i\in I(j)}\lambda_i^h x_{ij}^h/\mu_j^h$. Similarly, the form of constraint set (5-NP) arises from the fact that under non-preemptive priority, a high priority patient will not incur any wait at an EMS facility j if, upon arrival, it finds no other patients (high or low priority) ahead of it. The probability of this, by PASTA property, is given by $1-\sum_{c\in C}\Lambda_j^c/\mu_j^c = 1-\sum_{c\in C}\sum_{i\in I(j)}\lambda_i^c x_{ij}^c/\mu_j^c$. On the other hand, the analytical expression for the service level constraint (6) for the low priority patients is not known in absence of any closed-form expression for the waiting time distribution of low priority customers in a priority queue. Section 4 describes in detail the method to resolve this issue.

3.1 User Choice Environment

The model $[LAP]$ given by (1-7) is called a Directed Choice (DC) model, wherein the allocation of patients from a node i to an EMS facility j is dictated by a central authority. Such a model may be appropriate in environments in which patients from a given location are always brought by ambulances to a preassigned EMS facility, as dictated for example, by a central dispatching system. In this case, the patient has no control over which EMS facility to visit. However, a customer's choice of an EMS facility may not always be decided by the central authority but may be exercised by the user herself. The EMS location-allocation model in such an environment is referred as User Choice (UC) model. UC model requires explicitly specifying the decision model for users' choice of EMS facilities. For this, the most common assumption used in the literature is that users always choose the closest open service facility. Such an assumption is appropriate for walk-in patients or in environments where ambulances are required to bring patients to the closest EMS facility. The UC model can be represented by adding the Closest

Assignment Constraints (CAC). There are a variety of formulations of CAC available in the literature, the most widely cited among them having been proposed by Rojeski and ReVelle (1970). A version of their CAC, adapted to [LAP], is presented below.

$$x_{ij}^c \geq y_j - \sum_{l:d_{il}<d_{ij}} y_l \qquad \forall i \in I; j \in J(i); c \in C$$

The above CAC by Rojeski and ReVelle (1970) have the drawback that they become problematic in case of tied distances (travel time in the current problem) of two or more facilities with respect to the same user node. Several other formulations (Wagner and Falkson, 1975; Church and Cohon, 1976; Dobson and Karmarkar, 1987; Berman et al., 2006; Cánovas et al., 2007; Belotti et al., 2007; Marín, 2011) overcome this drawback. Espejo et al. (2012), based on a comparison of the different CAC formulations, identify the one proposed by Cánovas et al. (2007) as non-dominated by any other formulations in the literature. They also propose a new non-dominated CAC formulation. However, this new formulation can only be used in situations in which a predetermined number of service facilities needs to be opened. The CAC formulation by Cánovas et al. (2007), on the other hand, does not require this condition. Hence, we use their CAC formulation, a version of which adapted to [LAP] is presented below.

$$\sum_{l:d_{il}>d_{ij}} x_{il}^c + \sum_{l:d_{il}\leq d_{ij}, d_{kl}>d_{kj}} x_{kl}^c \leq 1 - y_j \qquad \forall i \in I; j \in J(i); c \in C \tag{8}$$

4 Solution Methodology

The absence of an analytical characterization of the service level constraint (6) for low priority patients makes $[LAP]$ challenging to solve. While the Laplace transform of the waiting time distribution $S_j^l(\tau^l)$, appearing in (6), and its first few moments are well known (Stephan, 1958), the distribution itself is somewhat complicated and requires numerical computation for the inverse Laplace transform, thereby preventing its analytical characterization (Jayaswal et al., 2010). There are approximations proposed in the literature for the waiting time distribution. However, they are very complex and often not sufficiently accurate (Abate and Whitt, 1997). Moreover, the choice of appropriate approximation to be used depends on Λ_j^h and Λ_j^l, which can only be determined endogenously, and are not known in advance in our model.

Although the exact form of $S^l_j(\tau^l)$ in (6) is unknown, it can be argued that it is concave in $(\Lambda^h_j, \Lambda^l_j)$. For a single priority (with homogeneous customers) M/M/1 queueing system, the cumulative distribution function (CDF) of customer waiting time (called service level in this paper) at EMS facility j is expressed as: $S^l_j(\tau^l) = P(W_j \leq \tau) = 1 - (\Lambda_j/\mu_j)e^{-(\mu_j - \Lambda_j)\tau}$, which is decreasing concave in Λ_j (it can be easily verified that its first two derivatives with respect to Λ_j are negative). So, in a queueing system with 2 customer classes, the CDF of waiting time of lower priority customers is expected to be decreasing concave in its own arrival rate. An increase in the arrival rate of higher priority customers is also expected to cause a decrease in the CDF of waiting time of lower priority customers since more high priority customers introduce more wait for the lower priority customers, and this increase is expected to be more rapid at higher arrival rates for high priority customers. Concavity of $S^l_j(\tau^l)$ in $(\Lambda^h_j, \Lambda^l_j)$ is further corroborated by the plot, determined numerically using the matrix geometric method, in Figure 1.

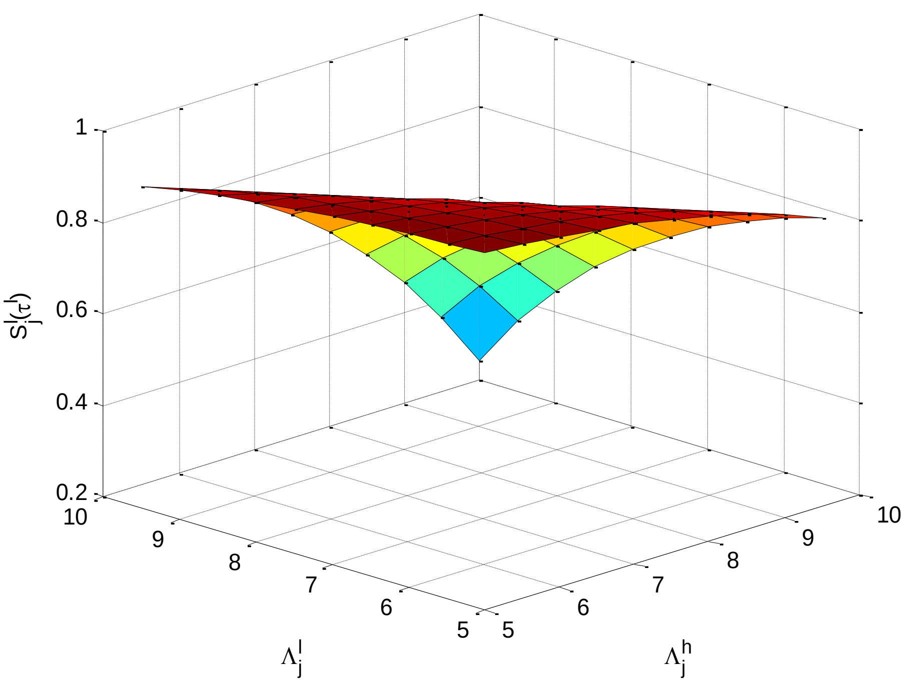

Figure 1: Service Level for Low Priority Patients at an EMS Facility vs. Demands for High Priority and Low Priority Patients under Preemptive Priority

We exploit the concavity of function $S^l_j(\tau^l)$ to approximate it arbitrarily closely by a set of tangent planes at various points $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$, $\forall\ p \in P$, as given below:

$$S_j^l(\tau^l) = \min_{p\in P}\left\{(S_j^l(\tau^l))^p + \left(\Lambda_j^h - (\Lambda_j^h)^p\right)\left(\frac{\partial(S_j^l(\tau^l))}{\partial\Lambda_j^h}\right)^p + \left(\Lambda_j^l - (\Lambda_j^l)^p\right)\left(\frac{\partial(S_j^l(\tau^l)}{\partial\Lambda_j^l}\right)^p\right\},$$

where $\left(S_j^l(\tau^l)\right)^p$ denotes the value of $S_j^l(\tau^l)$ at a fixed point $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$ and $\left(\frac{\partial(S_j^l(\tau^l))}{\partial\Lambda_j^h}\right)^p$ and $\left(\frac{\partial(S_j^l(\tau^l))}{\partial\Lambda_j^l}\right)^p$ are the partial gradients of $S_j^l(\tau^l)$ at $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$. Constraint set (6) can thus be replaced by the following set of linear constraints:

$$\left(S_j^l(\tau^l)\right)^p + \left(\Lambda_j^h - (\Lambda_j^h)^p\right)\left(\frac{\partial(S_j^l(\tau^l))}{\partial\Lambda_j^h}\right)^p + \left(\Lambda_j^l - (\Lambda_j^l)^p\right)\left(\frac{\partial(S_j^l(\tau^l)}{\partial\Lambda_j^l}\right)^p \geq \alpha^l \quad \forall p \in P \tag{9}$$

Substituting (9) in place of (6) results in a finite but a large number of constraints, which is amenable to cutting plane method (Kelley, 1960). We use the matrix geometric method to numerically evaluate $(S_j^l(\tau^l))^p$ at a given point $((\Lambda_j^h)^p$, $(\Lambda_j^l)^p$. The use of the matrix geometric method yields explicit recursive formula for the stationary wiating time distribution of low priority customers, which can provide significant computational improvements over the transform techniques. Moreover, it gives exact solutions, in contrast to simulation, which is another alternative method to evaluate S_j^l that at best gives point estimates. The matrix geometric method is also computationally efficient compared to simulation. This is important in solving (1 - 7) (for Directed Choice Environment) or (1 - 8) (for User Choice Environment), which requires repeated computation of $(S_j^l(\tau^l))^p$ for various open EMS facilities j at various solutions points p ($p \in P$). Once S_j^l is evaluated at a point $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$, its gradients are obtained using the *finite difference method* (described in Section 4.2). The gradients are used to generate cuts of the form (9), which are added iteratively in the cutting plane algorithm. The details of the cutting plane algorithm along with its computational performance are presented in Section 4.3.

4.1 Estimation of $S_j^l(\tau^l)$

In the following, we describe the matrix geometric method to evaluate the waiting time distribution of low priority patients, $S_j^l(\tau^l) = P(W_j^l \leq \tau^l)$, at a given point $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$ under preemptive priority. For the non-preemptive priority, the basic steps of the matrix geometric method remain the same as those for preemptive priority. So, we only briefly highlight the differences, relegating most of the details to appendix.

4.1.1 Estimation of $S_j^l(\tau^l)$ under Preemptive Priority

If we define $N_j^h(t)$ and $N_j^l(t)$ as state variables representing the number of high priority and low priority patients (in queue or in service) at EMS facility j at time t, then $\{\mathbf{N_j}(t)\} := \{N_j^l(t), N_j^h(t), t \geq 0\}$ is a continuous-time two-dimensional Markov chain with state space $\{\mathbf{n_j} = (n_j^l, n_j^h)\}$. In the context of two-dimensional Markov chains, we call n_j^l and n_j^h as the level and sub-level, respectively of the state space. As we will see below, $\{\mathbf{N_j}(t)\}$ is a *quasi birth-and-death* (QBD) process, permitting a matrix geometric solution for joint stationary distribution of $N_j^l(t)$ and $N_j^h(t)$. However, a general implementation of the matrix geometric method requires the number of sub-levels to be finite. For this, we assume $n_j^h \leq M$, where M is finite but large enough for the desired accuracy of our results. It is reasonable to assume a finite bound on the queue size of high priority patients since they are always served with preemptive priority.

In the Markov process $\{\mathbf{N_j}(t)\}$, a transition can occur only if a patient of either class arrives or is served at EMS facility j. The possible transitions are:

From	To	Rate	Condition
(n_j^l, n_j^h)	$(n_j^l, n_j^h + 1)$	Λ_j^h	for $n_j^l \geq 0$, $0 \leq n_j^h < M$
(n_j^l, n_j^h)	$(n_j^l + 1, n_j^h)$	Λ_j^l	for $n_j^l \geq 0$, $0 \leq n_j^h \leq M$
(n_j^l, n_j^h)	$(n_j^l, n_j^h - 1)$	μ_j^h	for $n_j^l \geq 0$, $0 \leq n_j^h \leq M$
(n_j^l, n_j^h)	$(n_j^l - 1, n_j^h)$	μ_j^l	for $n_j^l > 0$, $n_j^h = 0$

The transitions as described above result in the following infinitesimal generator Q:

$$
Q = \left(\begin{array}{c|cccc|cccc|cccc}
 & (0,0) & (0,1) & (0,...) & (0,M) & (1,0) & (1,1) & (1,...) & (1,M) & (2,0) & (2,1) & (2,...) & (2,M) \\
\hline
(0,0) & -\delta_1 & \Lambda_j^h & & & \Lambda_j^l & & & & & & & \\
(0,1) & \mu_j^h & -\delta_2 & \Lambda_j^h & & & \Lambda_j^l & & & & & & \\
(0,...) & & \mu_j^h & -\delta_2 & \Lambda_j^h & & & \Lambda_j^l & & & & & \\
(0,M) & & & \mu_j^h & -\delta_3 & & & & \Lambda_j^l & & & & \\
\hline
(1,0) & \mu_j^l & & & & -\delta_4 & \Lambda_j^h & & & \Lambda_j^l & & & \\
(1,1) & & & & & \mu_j^h & -\delta_2 & \Lambda_j^h & & & \Lambda_j^l & & \\
(1,...) & & & & & & \mu_j^h & -\delta_2 & \Lambda_j^h & & & \Lambda_j^l & \\
(1,M) & & & & & & & \mu_j^h & -\delta_3 & & & & \Lambda_j^l \\
\hline
(2,0) & & & & & \mu_j^l & & & & -\delta_4 & \Lambda_j^h & & \\
(2,1) & & & & & & & & & \mu_j^h & -\delta_2 & \Lambda_j^h & \\
(2,...) & & & & & & & & & & \mu_j^h & -\delta_2 & \Lambda_j^h \\
(2,M) & & & & & & & & & & & \mu_j^h & -\delta_3
\end{array}\right)
$$

where $\delta_1 = \Lambda_j^h + \Lambda_j^l$, $\delta_2 = \Lambda_j^h + \Lambda_j^l + \mu_j^h$, $\delta_3 = \Lambda_j^l + \mu_j^h$, and $\delta_4 = \Lambda_j^h + \Lambda_j^l + \mu_j^l$. Clearly, Q has a QBD structure, which upon grouping of all sub-levels for each level, can be represented as:

$$
Q = \begin{pmatrix}
L_0 & F & & \\
B & L & F & \\
 & B & L & F \\
 & & \ddots & \ddots & \ddots
\end{pmatrix}
$$

where L_0, F, L, B are square matrices of size $M+1$. This allows us to develop a matrix geometric solution for the joint distribution of the number of patients of each class at EMS facility j.

We denote $\mathbf{x} = [\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_k}, \ldots, \ldots]$ as the stationary probability vector of $\{\mathbf{N_j}(t)\}$, where $\mathbf{x}_k = [x_{k0}, x_{k1}, \ldots, x_{kM}]$ is the stationary probability of different sub-levels in level k ($n_j^l = k$). $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form by the following standard relations (Neuts, 1981):

$$
\mathbf{x}Q = \mathbf{0}; \quad \mathbf{x}_{k+1} = \mathbf{x}_k R \quad \forall k \geq 0
$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$F + RL + R^2B = \mathbf{0}$$

The matrix R can be computed using well known methods (Latouche and Ramaswai, 1999). A simple iterative procedure often used is:

$$R(0) = 0 \;; \quad R(n+1) = -\left[F + R^2(n)B\right] L^{-1}$$

The probabilities $\mathbf{x}_0$ are determined using:

$$\mathbf{x}_0(L_0 + RB) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{k=0}^{\infty} \mathbf{x}_k \mathbf{1} = \mathbf{x}_0 (I - R)^{-1} \mathbf{1} = 1$$

where $\mathbf{1}$ is a column vector of ones of size $M + 1$.

The waiting time W_j^l of a low priority patient at EMS facility j is the time between its arrival to the facility j till it first enters into service at that facility. It is difficult to characterize the stationary distribution $S_j^l(\cdot)$ of W_j^l. However, Ramaswami and Lucantoni (1985) present an efficient algorithm to numerically compute the complimentary distribution of waiting times in QBD processes. Jayaswal et al. (2011) adapt their algorithm to compute the sojourn time (waiting time plus the time in service) distribution of low priority customers, which we adopt (with modification for waiting time in queue) in this paper.

Consider a tagged low priority patient entering facility j. We now redefine level of the system as the number of low priority patients observed by the tagged patient upon its arrival at facility j, instead of the total number of low priority patients at facility j as described in section above. The time spent by the tagged patient in waiting at facility j depends on the number of patients of either class already present at facility j ahead of it, and also on the number of subsequent arrivals of high priority patients before it (the tagged patient) enters into service. All subsequent arrivals of low priority patients to facility j, however, have no influence on the waiting time of the tagged patient. We, therefore, set $\Lambda_j^l = 0$ for the purpose of computing $S_j^l(\cdot)$.

Further, if we set all the transition rates out of state $(0,0)$ to 0, then state $(0,0)$ becomes an absorbing state, and the waiting time of the tagged patient is simply the time until absorption in this modified Markov process $\{\tilde{\mathbf{N}}_{\mathbf{j}}(t)\}$ with the infinitesimal generator $\tilde{Q}$ as given below:

$$\tilde{Q} = \left(\begin{array}{c|c|cccc|cccc|cccc}
 & 0^* & (0,1) & (0,...) & (0,M) & (1,0) & (1,1) & (1,...) & (1,M) & (2,0) & (2,1) & (2,...) & (2,M) \\
\hline
0^* & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
(0,1) & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h & & & & & & & & & \\
(0,...) & & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h & & & & & & & & \\
(0,M) & & & \mu_j^h & -\tilde{\delta}_3 & & & & & & & & \\
\hline
(1,0) & \mu_j^l & & & & -\tilde{\delta}_4 & \Lambda_j^h & & & & & & \\
(1,1) & & & & & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h & & & & & \\
(1,...) & & & & & & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h & & & & \\
(1,M) & & & & & & & \mu_j^h & -\tilde{\delta}_3 & & & & \\
\hline
(2,0) & & & & & \mu_j^l & & & & -\tilde{\delta}_4 & \Lambda_j^h & & \\
(2,1) & & & & & & & & & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h & \\
(2,...) & & & & & & & & & & \mu_j^h & -\tilde{\delta}_2 & \Lambda_j^h \\
(2,M) & & & & & & & & & & & \mu_j^h & -\tilde{\delta}_3
\end{array}\right)$$

where $\tilde{\delta_2} = \Lambda_j^h + \mu_j^h$, $\tilde{\delta_3} = \mu_j^h$, and $\tilde{\delta_4} = \Lambda_j^h + \mu_j^l$. State $(0,0)$ in $\tilde{Q}$ is now indicated using a special notation 0^* to emphasize that it is an absorbing state. $\tilde{Q}$, upon grouping of all sub-levels for each level, can be represented as:

$$\tilde{Q} = \left(\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & \cdots \\
\hline
b_0 & \tilde{L}_0 & 0 & & & \\
b_1 & 0 & \tilde{L} & 0 & & \\
0 & & B & \tilde{L} & 0 & \\
\vdots & & & \ddots & \ddots & \ddots
\end{array}\right)$$

where, $\tilde{L_0}$ is now a square matrix of size M due to the removal of the state $(0,0)$. For the same reason, b_0 is a column vector of size M.

The distribution $S_j^l(y)$ of the time spent by a low priority patient at facility j can be expressed as:

$$\begin{aligned} S_j^l(y) &= 1 - \overline{S_j^l}(y) && \text{for } y > 0 \\ &= x_{00} && \text{for } y = 0 \end{aligned}$$

where $\overline{S_j^l}(y)$ is the stationary probability that a low priority patient spends more than y units of time at facility j. $S_j^l(y = 0) = x_{00}$ accounts for the possibility for the tagged patient to find the system empty, i.e., in the absorbing state 0^*, upon its arrival to facility j, in which case its waiting time is 0. Let $\overline{S_{jk}^l(y)}$ denote the conditional probability that the tagged patient, which finds k low priority patients ahead of it (i.e., level k) upon arrival at facility j spends a time exceeding y before entering into service. The probability that the tagged patient, upon arrival at facility j, finds k low priority patients ahead of it is given, using the PASTA property, by $\mathbf{x}_k = \mathbf{x}_0 R^k$. Using the law of total probability, $\overline{S_j^l}(y)$, in turn, can be expressed as:

$$\overline{S_j^l}(y) = \tilde{\mathbf{x}}_{\mathbf{0}} \overline{S_{j0}^l}(y) + \sum_{k=1}^{\infty} \mathbf{x}_{\mathbf{k}} \overline{S_{jk}^l}(y) \tag{10}$$

where $\tilde{\mathbf{x}}_{\mathbf{0}} = [x_{01}, \ldots, x_{0M}]$. In other words, $\tilde{\mathbf{x}}_{\mathbf{0}}$ is the probability of the system being in level 0 (corresponding to 0 low priority patients), as described above, with state $(0, 0)$ (corresponding to empty system) removed from the level.

Each of the terms in (10) can be computed more conveniently by uniformizing the Markov process $\{\tilde{\mathbf{N}}_{\mathbf{j}}(t)\}$ with a Poisson process with rate γ, where

$$\gamma = \max_{0 \le m \le M} |(\tilde{L})| = \max\{\tilde{\delta}_2, \tilde{\delta}_3, \tilde{\delta}_4\}$$

so that the rate matrix $\tilde{Q}$ is transformed into the discrete-time probability matrix:

$$\hat{Q} = \frac{1}{\gamma}\tilde{Q} + I = \left(\begin{array}{c|ccccc} 1 & 0 & 0 & 0 & 0 & \cdots \\ \hline \hat{b}_0 & \hat{L}_0 & 0 & & & \\ \hat{b}_1 & 0 & \hat{L} & 0 & & \\ 0 & & \hat{B} & \hat{L} & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\hat{B} = \frac{B}{\gamma}$, $\hat{L}_0 = \frac{\tilde{L}_0}{\gamma} + I$, $\hat{L} = \frac{\tilde{L}}{\gamma} + I$, $\hat{b}_0 = \frac{b_0}{\gamma}$, $\hat{b}_1 = \frac{b_1}{\gamma}$, and I is an identity matrix of appropriate dimension. In this uniformized process, points of a Poisson process are generated with a rate γ, and transitions occur at these epochs only. The probabilities that a transition at such an epoch only involves a change in sub-levels (i.e., the number of high priority patients) and no change in levels (i.e., the number of low priority patients) are given by the elements of $\hat{L}_0$ for level $k = 0$, and by the elements of $\hat{L}$ for level $k \geq 1$. On the other hand, the probabilities that a transition at such an epoch involves a decrease in level not leading to absorption are given by the elements of $\hat{B}$ for level $k \geq 2$. Such probabilities are all equal to 0 for level $k = 1$, as clear from $\hat{Q}$ matrix shown above.

The probability that n Poisson events are generated in time y is given by $e^{-\gamma y}\frac{(\gamma y)^n}{n!}$. Suppose the tagged patient finds $k > 0$ low priority patients ahead of it. Then, for its waiting time at facility j to exceed y, at most $k - 1$ of the n generated Poisson points may correspond to transitions to lower levels (i.e., service completions of low priority patients). We use this argument to compute each of the terms in (10) as follows.

$$\tilde{\mathbf{x}}_{\mathbf{0}}\overline{S^l_{j0}}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\tilde{\mathbf{x}}_{\mathbf{0}} G^{(n)}_{00}\mathbf{1} \tag{11}$$

$$\mathbf{x}_{\mathbf{k}}\overline{S^l_{jk}}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\mathbf{x}_{\mathbf{k}} \sum_{v=0}^{k-1} G^{(n)}_{v}\mathbf{1} \qquad \text{for } k \geq 1 \tag{12}$$

where the entries of the matrix $G^{(n)}_{00}$ represent the conditional probabilities that the process, given that it starts in level 0, remains in level 0 after n transitions in the discrete-time Markov process with rate matrix $\hat{Q}$. $G^{(n)}_v$ is a matrix such that its entries are the conditional probabilities, given that the system has made n transitions in the discrete-time Markov process with rate matrix $\hat{Q}$, that v of those transitions correspond to lower levels. Matrices $G^{(n)}_{00}$ and $G^{(n)}_v$ can be computed recursively as:

$$G^{(n)}_{00} = G^{(n-1)}_{00}\hat{L}_0; \qquad G^{(0)}_{00} = I. \tag{13}$$

$$G^{(n)}_{v} = G^{(n-1)}_{v-1}\hat{B} + G^{(n-1)}_{v}\hat{L} \tag{14}$$

Using (12):

$$\sum_{k=1}^{\infty}\mathbf{x_k}\overline{S^l_{jk}}(y) = \sum_{k=1}^{\infty}\sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\mathbf{x_k}\sum_{v=0}^{k-1} G_v^{(n)}\mathbf{1}$$
$$= \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\mathbf{x_0}\sum_{k=1}^{\infty} R^k \sum_{v=0}^{k-1} G_v^{(n)}\mathbf{1} \tag{15}$$

Now,

$$\sum_{k=1}^{\infty} R^k \sum_{v=0}^{k-1} G_v^{(n)}\mathbf{1}$$
$$= \sum_{k=1}^{n+1} R^k \sum_{v=0}^{k-1} G_v^{(n)}\mathbf{1} + \sum_{k=n+2}^{\infty} R^k \sum_{v=0}^{n} G_v^{(n)}\mathbf{1} \qquad \left(\text{since} \quad G_v^{(n)} = 0 \quad \text{for} \quad v > n\right)$$
$$= \sum_{v=0}^{n}\sum_{k=v+1}^{n+1} R^k G_v^{(n)}\mathbf{1} + (I-R)^{-1}R^{n+2}\mathbf{1} \qquad \left(\text{since} \quad \sum_{v=0}^{n} G_v^{(n)}\mathbf{1} = \mathbf{1}\right)$$
$$= \sum_{v=0}^{n}(I-R)^{-1}(R^{v+1}-R^{n+2})G_v^{(n)}\mathbf{1} + (I-R)^{-1}R^{n+2}\mathbf{1}$$
$$= \sum_{v=0}^{n}(I-R)^{-1}R^{v+1}G_v^{(n)}\mathbf{1} \qquad \left(\text{since} \quad \sum_{v=0}^{n} G_v^{(n)}\mathbf{1} = \mathbf{1}\right)$$
$$= (I-R)^{-1}RH^{(n)}\mathbf{1} \tag{16}$$

where,

$$H^{(n)} = \sum_{v=0}^{n} R^v G_v^{(n)} \qquad \text{for } n \geq 0. \tag{17}$$

Susbtituting (16) in (15) gives:

$$\sum_{k=1}^{\infty}\mathbf{x_k}\overline{S^l_{jk}}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\mathbf{x}_0(I-R)^{-1}RH^{(n)}\mathbf{1} \tag{18}$$

Using (17) in (14) gives us the following recursive formula to compute $H^{(n)}$:

$$\begin{aligned} H^{(n)} &= \sum_{v=0}^{n} R^v G_v^{(n)} \\ &= \sum_{v=0}^{n} R^v \left(G_{v-1}^{(n-1)} \hat{B} + G_v^{(n-1)} \hat{L} \right) \\ &= R \sum_{v=1}^{n} R^{v-1} G_{v-1}^{(n-1)} \hat{B} + \sum_{v=0}^{n-1} R^v G_v^{(n-1)} \hat{L} \qquad \left(\text{since} \quad G_n^{(n-1)} = 0\right) \\ &= R H^{(n-1)} \hat{B} + H^{(n-1)} \hat{L}; \quad H^{(0)} = R^0 G_0^{(0)} = I \qquad \left(\text{since} \quad R^0 = I \quad \text{and} \quad G_0^{(0)} = I\right) \end{aligned}$$

Using (11) and (18) in (10), we get:

$$\overline{S_j^l}(y) = \sum_{n=0}^{\infty} e^{-\gamma y} \frac{(\gamma y)^n}{n!} \left\{ \tilde{\mathbf{x}}_{\mathbf{0}} G_{00}^{(n)} \mathbf{1} + \mathbf{x}_0 (I - R)^{-1} R H^{(n)} \mathbf{1} \right\} \tag{19}$$

Therefore, for given arrival rates $\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)$ at facility j, $S_j^l(\tau^l) = 1 - \overline{S_j^l}(\tau^l)$ in (9) can be computed using (19).

4.1.2 Estimation of $S_j^l(\tau^l)$ under Non-preemptive Priority

Under non-preemptive priority, to completely describe the state of the system, one needs to also specify the class (high or low priority) of customer in service when there are both classes of customers in the system. For that, let $Z_j(t)$ represent the class of patient being served, when there are both classes of patients at an EMS facility, at time t. Then $\{\mathbf{N_j}(t)\} := \{N_j^l(t), Z_j(t), N_j^h(t), t \geq 0\}$ is a continuous-time three-dimensional Markov chain with state space $\{\mathbf{n_j} = (n_j^l, z_j, n_j^h)\}$ and possible transitions among the states as given in Figure 2. We group the states and define level k as: $\{(n_j^l, z_j, n_j^h) | n_j^l = k, z_j \in \{h, c\}, 0 \leq n_j^h \leq M\}$. Within level k, any feasible combination of $\{(z_j, n_j^h)\}$ is called a sub-level. If the sub-levels with a level k are arranged lexicographically such that $(k, h, n_j^h) < (k, l, n_j^h)$, then the above transition diagram results in the following infinitesimal generator Q: Clearly, Q has a QBD

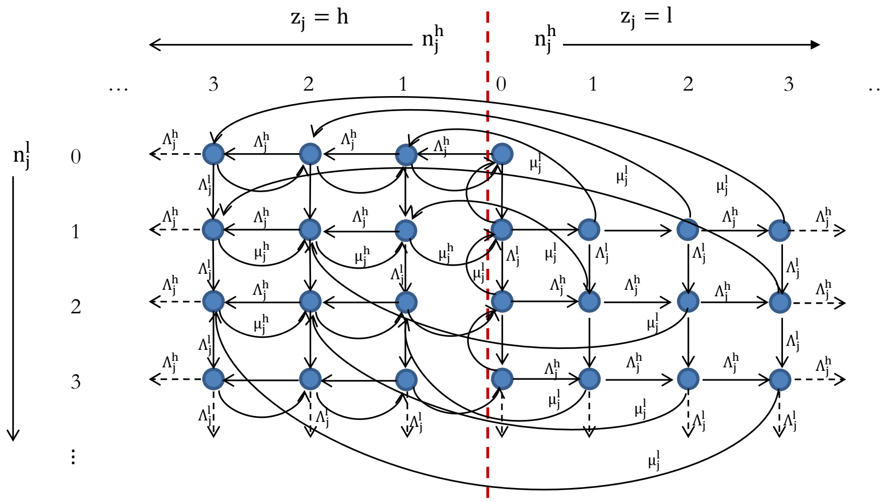

Figure 2: Transition Diagram for Non-preemptive Priority Queue

structure, which upon grouping of all sub-levels for each level, can be represented as:

$$Q = \begin{pmatrix} L_0 & F_0 & & & \\ B_0 & L & F & & \\ & B & L & F & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where B, L, F are square matrices of size $2M+1$. L_0 is a square matrix of size $M+1$, while B_0, F_0 are of sizes $(2M+1)\times(M+1)$ and $(M+1)\times(2M+1)$, respectively. These matrices can be easily constructed using the transition rates described above, and are provided in Appendix 1.

We denote $\mathbf{x} = [\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_k}, \ldots, \ldots]$ as the stationary probability vector of $\{\mathbf{N_j}(t)\}$, where $\mathbf{x}_0 = [x_{00}, x_{01}, \ldots, x_{0M}]$ and $\mathbf{x}_k = [x_{k0}, x_{kh1}, \ldots, x_{khM}, x_{kl1}, \ldots, x_{klM}]$ for $k \geq 1$ are the stationary probabilities of different sub-levels in level k ($n_j^l = k$). $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form by the following standard relations:

$$\mathbf{x}Q = \mathbf{0}; \quad \mathbf{x}_{k+1} = \mathbf{x}_k R \quad \forall k \geq 1$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$F + RL + R^2 B = \mathbf{0}$$

The probabilities $\mathbf{x}_0$ are determined as the solution to the following system of equations:

$$\mathbf{x}_0 L_0 + \mathbf{x}_1 B_0 = \mathbf{0}; \quad \mathbf{x}_0 F_0 + \mathbf{x}_1 (L + RB) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{k=0}^{\infty} \mathbf{x}_k \mathbf{1} = \mathbf{x}_0 \mathbf{1} + \mathbf{x}_1 (I - R)^{-1} \mathbf{1} = 1$$

The distribution $S_j^l(y)$ of the time spent by a low priority patient at facility j under non-preemptive priority can also be expressed as:

$$\begin{aligned} S_j^l(y) &= 1 - \overline{S_j^l}(y) && \text{for } y > 0 \\ &= x_{00} && \text{for } y = 0 \end{aligned}$$

where, the expression for $S_j^l(y)$ under non-preemptive priority can be derived using the same arguments as for preemptive priority described in section 4.1.1. The final expression for $S_j^l(y)$ under non-preemptive priority is given by (20).

$$\overline{S_j^l}(y) = \sum_{n=0}^{\infty} e^{-\gamma y} \frac{(\gamma y)^n}{n!} \left\{ \tilde{\mathbf{x}}_{\mathbf{0}} G_{00}^{(n)} \mathbf{1} + \mathbf{x}_1 (I - R)^{-1} H^{(n)} \mathbf{1} \right\} \tag{20}$$

where, where $\tilde{\mathbf{x}}_{\mathbf{0}} = [x_{01}, \ldots, x_{0M}]$.

4.2 Estimation of the Gradient of $S_j^l(\tau^l)$

There are several methods available in the literature to compute the partial gradients of $S_j^l(\tau^l)$. We use a *finite difference* method as it is probably the simplest and most intuitive, and can be easily explained (Atlason et al., 2004). Finite difference method can further be employed either as the central difference, forward difference or the backward difference. Using the *central* *difference* method, we compute gradients as:

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^h} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p + d\Lambda^h, (\Lambda_j^l)^p\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p - d\Lambda^h, (\Lambda_j^l)^p\right)}}{2d\Lambda^h}$$

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^l} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p + d\Lambda^l\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p - d\Lambda^l\right)}}{2d\Lambda^l}$$

where $d\Lambda^h$, $d\Lambda^l$ (referred to as step sizes) are infinitesimal changes in the respective variables. However, when $(\Lambda_j^h)^p < d\Lambda^h$ or $(\Lambda_j^l)^p < d\Lambda^l$, then the corresponding gradient is estimated using the forward difference method as:

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^h} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p + d\Lambda^h, (\Lambda_j^l)^p\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)}}{d\Lambda^h}$$

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^l} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p + d\Lambda^l\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)}}{d\Lambda^l}$$

On the other hand, when $(\Lambda_j^h)^p \geq \mu_j^h - d\Lambda^h$ or $(\Lambda_j^l)^p \geq \mu_j^l - d\Lambda^l$, then the corresponding gradient is estimated using the backward difference method as:

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^h} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p - d\Lambda^h, (\Lambda_j^l)^p\right)}}{d\Lambda^h}$$

$$\frac{\partial \left(S_j^l(\tau^l)\right)^p}{\partial \Lambda_j^l} = \frac{\left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p\right)} - \left(S_j^l(\tau^l)\right)^{\left((\Lambda_j^h)^p, (\Lambda_j^l)^p - d\Lambda^l\right)}}{2d\Lambda^l}$$

4.3 The Cutting Plane Algorithm

In this section, we describe the cutting plane algorithm to solve $[LAP]$. The algorithm fits the framework of Kelley's cutting plane method (Kelley, 1960). It differs from the traditional description of the algorithm in that we use the matrix geometric method to generate the cuts and evaluate the function values instead of having an algebraic form for the function and using analytically determined gradients to generate the cuts. The steps of the algorithm are outlined below:

The algorithm starts with an empty constraint set (9), and obtain an initial solution resulting in $\left(\left(\Lambda_j^h\right)^0, \left(\Lambda_j^l\right)^0\right)$ at EMS facility $j \in \left\{J : (y_j)^0 = 1\right\}$. We use the matrix geometric method to

Algorithm 1 Cutting Plane Algorithm

1: $p \leftarrow 0$.
2: **repeat**
3: Solve $[LAP]$ ((1)-(4), (5-P), (7) for preepmtive priority under DC; (1)-(4), (5-P), (7), (8) for preepmtive priority under UC; (1)-(4), (5-NP), (7) for non-preepmtive priority under DC; (1)-(4), (5-NP), (7), (8) for non-preepmtive priority under UC) to obtain $\left(x_{ij}^{c}\right)^{p}$ $\forall c \in \{h, l\}$ and $(y_j)^p \quad \forall j \in J$.
4: Obtain $\left(\Lambda_j^h\right)^p = \sum_{i \in I} \lambda_j^h \left(x_{ij}^h\right)^p$ and $\left(\Lambda_j^l\right)^p = \sum_{i \in I} \lambda_j^l \left(x_{ij}^l\right)^p \quad \forall j \in \{J : (y_j)^p = 1\}$.
5: Obtain $\left(S_j^l(\tau^l)\right)^p$ using (19) for preemptive priority or using (20) for non-preemptive priority $\forall j \in \{J : (y_j)^p = 1\}$.
6: **if** $\left(S_j^l(\tau^l)\right)^p \geq \alpha^l \ \forall j \in \{J : (y_j)^p = 1\}$ **then**
7: Stop.
8: **else**
9: Add to $[LAP]$ cuts of the form (9) $\forall j \in \left\{J : (y_j)^p = 1 : \left(S_j^l(\tau^l)\right)^p < \alpha^l\right\}$.
10: $p \leftarrow p + 1$.
11: **end if**
12: **until** $\left(S_j^l(\tau^l)\right)^p < \alpha^l$ for any $j \in \{J : (y_j)^p = 1\}$.

compute the distribution $\left(S_j^l(\tau^l)\right)^{\left(\left(\Lambda_j^h\right)^0, \left(\Lambda_j^l\right)^0\right)}$ of W_j^l. If $\left(S_j^l(\tau^l)\right)^{\left(\left(\Lambda_j^h\right)^0, \left(\Lambda_j^l\right)^0\right)}$ meets the delivery time reliability constraint $\alpha^l \ \forall j \in \left\{J : (y_j)^0 = 1\right\}$, we stop with an optimal solution to $[LAP]$, else we add to (9) linear constraints generated using the finite difference method. The new cuts eliminate the current solution but do not eliminate any feasible solution to $[LAP]$. This procedure repeats until the service level constraint for low priority patients is satisfied at all EMS facilities within a sufficiently small tolerance limit ϵ such that $\left|S_j^l(\tau^l) - \alpha\right| \leq \epsilon$. The method has been proved to converge (Atlason et al., 2004).

5 Computational Results and Discussion

In this section, we report the performance of our solution method. Algorithm 1 is coded in Visual C++, while the model $[LAP]$ in step 3 of the algorithm is solved using IBM CPLEX 12.4. All the experiments are performed on a Pentium i5-3470, 3.20GHz, 64-bit PC with 8GB RAM. The data used in this study are presented in section 5.1. In section 5.2, we present an illustrative example to demonstrate the steps of Algorithm 1, as described in section 4. Results of our extensive computational experiments are presented in section 5.3.

5.1 Data

For the illustrative example and computational experiments, presented in sections 5.2 and 5.3, we use the 5-month period census tract level data for Austin, Texas, USA as reported by Daskin and Stern (1981) and Daskin (1982). Figure 3 depicts zones (33 census tracts) on the map of Austin, Texas (adopted from Daskin (1982)). Table A1 (in appendix) shows the EMS service call data from the 33 zones, collected over a 5-month period. We assume that the EMS service calls (demands) from a given zone $i \in I$ arise according to stationary Poisson process at an hourly rate λ_i obtained by dividing the 5-month service call data by $3,600$ $(= 5 \times 30 \times 24)$, which is the number of hours in a 5-month period. In the problem described in this paper, patients are triaged, as described in section 3, as resuscitation/high priority (denoted by h) that require immediate access, or less urgent/low priority (denoted by l), which subsumes all the remaining acuity levels. In absence of acuity level demand data, we assume fixed proportions $f_i^h = f^h \in (0,1)$ and $f_i^l = 1 - f^h \ \forall i \in I$ of the demand from any zone arise from high priority and low priority patients, respectively, such that $\lambda_i^h = f_i^h \lambda_i$ and $\lambda_i^l = f_i^l \lambda_i$. Each user zone is also a candidate site for EMS facility, such that $J = I$.

The inter-zonal travel times are given by the travel time matrix shown in Table A2 (in appendix). The travel time matrix presents only the travel times from zone number i to zone number $j : j \geq i$; those from i to $j : j < i$ are implied from the symmetry of the matrix. We define the coverage radius $R = 10$ minutes, same as used by Daskin (1982), in all our experiments. The travel cost is set to TC = \$ 1 per patient minute, and the (amortized) cost of opening and operating an EMS facility at location j as $FC_j =$ \$ 100. Note that the objective function of model $[LAP]$ has following two components: the first component minimizes the total cost of EMS facilities to be located; and the second component minimizes the total travel cost of all the patients to their allocated EMS facilities. By ensuring a sufficiently high coefficient for the first component, compared to the second one, the problem always attempts to locate the minimum number of EMS facilities before seeking to minimize the total time travelled by all the patients in the network to the respective EMS facilities that they patronize.

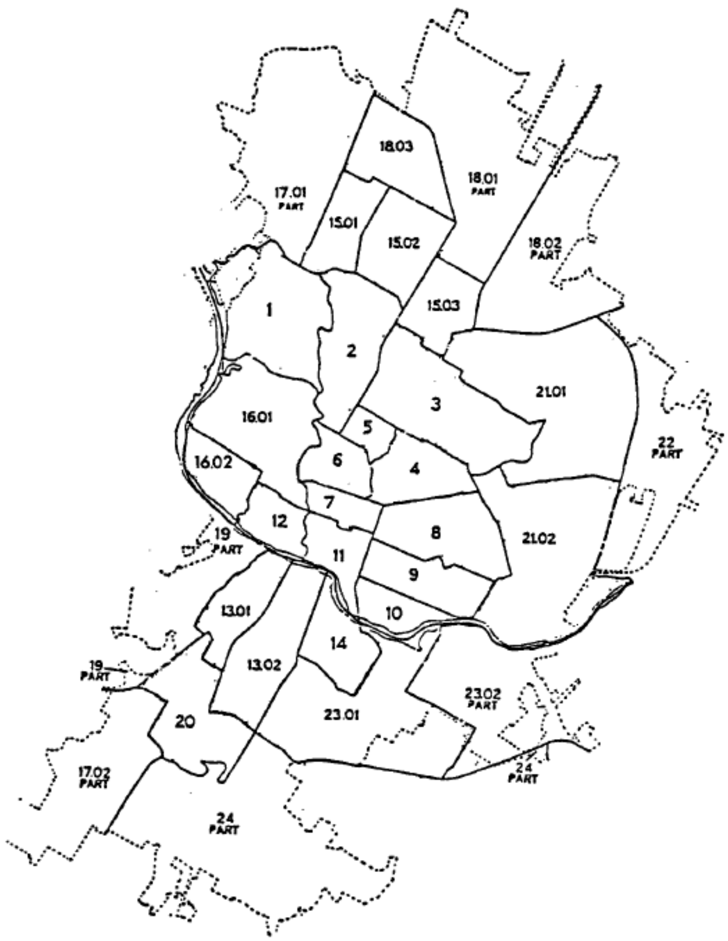

Figure 3: Census Tracts of Austin, Texas (Adopted from Daskin (1982))

5.2 Illustrative Example

We illustrate the steps of Algorithm 1 for the preemptive priority under DC version of $[LAP]$ using an example generated from the data as described in section 5.1. For the purpose of illustration, we fix the proportion f_i^h of the total service calls arising from any zone that are triaged as high priority at 1% $\forall i \in I$. This closely matches the observation made in the 2010 annual report of the office of the Auditor General of Ontario[3], which indicates that only 0.6% of the total emergency-department visits in the hospitals in Ontario constituted resuscitation cases. The service rates for both high and low priority patients at any EMS facility are fixed as $\mu_j^h = 2$, $\mu_j^l = 2$ per hour $\forall j \in J$. The service level requirements for high and low priority patients are specified as follows:

- 98% of the high priority patients arriving at any EMS facility should be provided emergency care immediately after triage, i.e., $S_j^h(\tau^h = 0) = P(W_j^h \leq 0) \geq \alpha^h = 0.98 \ \forall j \in J$.
- 90% of the low priority patients arriving at any EMS facility should be provided emergency care within 15 minutes after triage, i.e., $S_j^l(\tau^l = 15) = P(W_j^l \leq 15) \geq \alpha^l = 0.90 \ \forall j \in J$.

Algorithm 1 solves the above problem in 15 seconds using 6 iterations. The location-allocation decisions, and the resulting service levels achieved in each iteration are indicated in Figures 4 and 5. As discussed above, solving $[LAP]$ is challenging due to absence of an analytical expression for $S_j^l(\tau^l)$ appearing in (6). To overcome this, we exploit the concavity of $S_j^l(\tau^l)$, as argued and also verified using matrix geometric method in section 4, to approximate it using linear constraints of the type (9), which are dynamically generated as they are needed.

Algorithm 1 starts with the constraint set (9) being empty (corresponding to $p = 0$ in step 1). This results in 4 EMS facilities getting opened in zones $2, 8, 23$ and 31, and a total travel time (TT) in patient minutes per hour of 6.003. The allocations of user nodes to these facilities are shown in Iteration 1 of Figure 4. This results in an achieved service level of only 87.2% and 89.9% for low priority customers at EMS facilities located in zones 8 and 31, respectively. This can be overcome by reducing the traffic intensity seen by the EMS facility in zones 8 and 31. For this, cuts $0.372\Lambda_8^h + 0.366\Lambda_8^l \leq 0.112$ and $0.358\Lambda_{31}^h + 0.353\Lambda_{31}^l \leq 0.108$ are generated, using the method described in section 4.2, and added to the model $[LAP]$. The resulting

[3] http://www.auditor.on.ca/en/reports_en/en10/305en10.pdf

model is resolved (corresponding to $p = 1$), which results in the EMS facility in zone 8 getting replaced by another one in zone 4 with re-allocations of patients (indicated by blue colored lines). The re-allocations of users results in an increase in TT from 6.003 to 6.212. This also pushes the service levels for low priority patients at EMS facility in zone 31 above the required 90% mark but pulls the same at the new EMS facility in zone 4 down to 85.3%. To satisfy the service level requirement at the EMS facility in zone 4, the algorithm now adds the cut $0.382\Lambda_4^h + 0.375\Lambda_4^l \leq 0.116$. The resulting model is again resolved (corresponding to $p = 2$), and the process repeats until the service level is at least 90% at all the open EMS facilities. The location and allocation of EMS facilities at each of the 6 iterations of the algorithm are shown in Figures 4 and 5.

5.3 Computational Results

In Table 2, we report the performance of our proposed solution method for a range of problem parameters for preemptive priority under UC. The service level requirement for the high priority (resuscitation) patients is set, according to CTAS guidelines, as $\alpha^h = 98\%$ of the patients be served immediately after triage (refer Table 1). For the lower priority class, we vary the service requirement α^l in the set $\{80\%, 85\%, 90\%, 95\%\}$ according to CTAS guidelines. The patient mix is varied using the the following values: $f_i^h = 0.5\%,\ 1\%,\ 5\%\ \forall i \in I$. The service rates for both high and low priority patients at any EMS facility are fixed as $\mu_j^h = 2$, $\mu_j^l = 2$ per hour $\forall j \in J$. The table reports the number of facilities opened (NF), total travel time in patient minutes per hour (TT), and the locations of facilities (Facility) and the service level achieved at opened facilities ($[S^h, S^l]$ in %), the comptational time in seconds (CPU) and the iterations (Iter.) taken by the algorithm.

The results suggest that the algorithm solves most of the problem instances within a few seconds. Specifically, when the service level requirement for low priority customers is not very stringent, for example $S^l(\tau^l = 60) \geq 80\%$ or $S^l(\tau^l = 120) \geq 80\%$, then the algorithm solves the problem in a second, requiring only one iteration. This is so because for relatively low service level requirements for low priority patients, the EMS facility locations and their allocations implied by the service level requirement for the high priority patients are sufficient to also guarantee the service level requirement for the low priority patients. However, this is no longer true as the service level requirement for low priority becomes tighter. In such a case, the cutting plane algorithm gets invoked, requiring multiple iterations to solve the problem. This is specially evident for $f^h = 5\%$, $f^l = 95\%$, $S^l(\tau^l = 15) \geq 95\%$ in which case the algorithm takes 15 iterations.

5.3.1 Directed Choice versus User Choice

An interesting question that arises in the context of EMS facility location problem is whether the users are, on average, better off when the system lets them decide which service facility to seek service from compared to the case when the system decides it for them (Boffey et al., 2007). For this, in Table 3, we present a comparison of the total minutes travelled per hour (TT) in the network under DC versus UC for both preemptive and non-preemptive priority. A lower value for TT under UC compared to that under DC indicates that users are better off under the former. Our results suggest that often, it makes no difference to the users whether the system decides for them or let them decide which EMS facility to patronize, as indicated by a zero value for the % change in TT between DC and UC. This is in agreement with the observation made by Aboolian et al. (2012), although in a slightly different context (see section 2 for their problem context). However, users' utility, on average, (as captured by TT) is not always the same under DC and UC. For example, for $\mu^h = 2$, $\mu^l = 2$, $f^h = 0.5\%$, $f^l = 99.5\%$, $\alpha^h = 98\%$, $\alpha^l = 95\%$ under preemptive priority, users are, on average, better off under UC than under DC, as indicated by a positive value for % change in TT $= 100\frac{TT(DC)-TT(UC)}{TT(DC)}$. This is because the Closest Assignment Constraints (CAC) make the optimal solution obtained for DC infeasible for UC. Hence, to satisfy CAC, in addition to the service level constraints, the model is forced to choose a solution that is sub-optimal under DC, which in this instance turns out to be one with 6, instead of 5 under DC, EMS facilities. An extra EMS facility under UC reduces the total patient minutes travelled by all patients (TT), resulting in a positive % change in TT.

What is surprising is the observation that users, on average, can even be worse off deciding by themselves which service facility to patronize, as indicated by a negative value for % change in TT for $\mu^h = 3$, $\mu^l = 2$, $f^h = 5\%$, $f^l = 95\%$, $\alpha^h = 98\%$, $\alpha^l = 95\%$ under preemptive priority. This happens again because if each user zone is assigned to its closest EMS facility among those opened under DC, then this violates either or both of the service level constraints (5) and (6) at some of the open facilities. Hence, to satisfy CAC, in addition to the service level constraints, the model under UC is forced to choose a different set of 4 EMS facilities, which is sub-ptimal under DC. This results in a negative value for % change in TT.

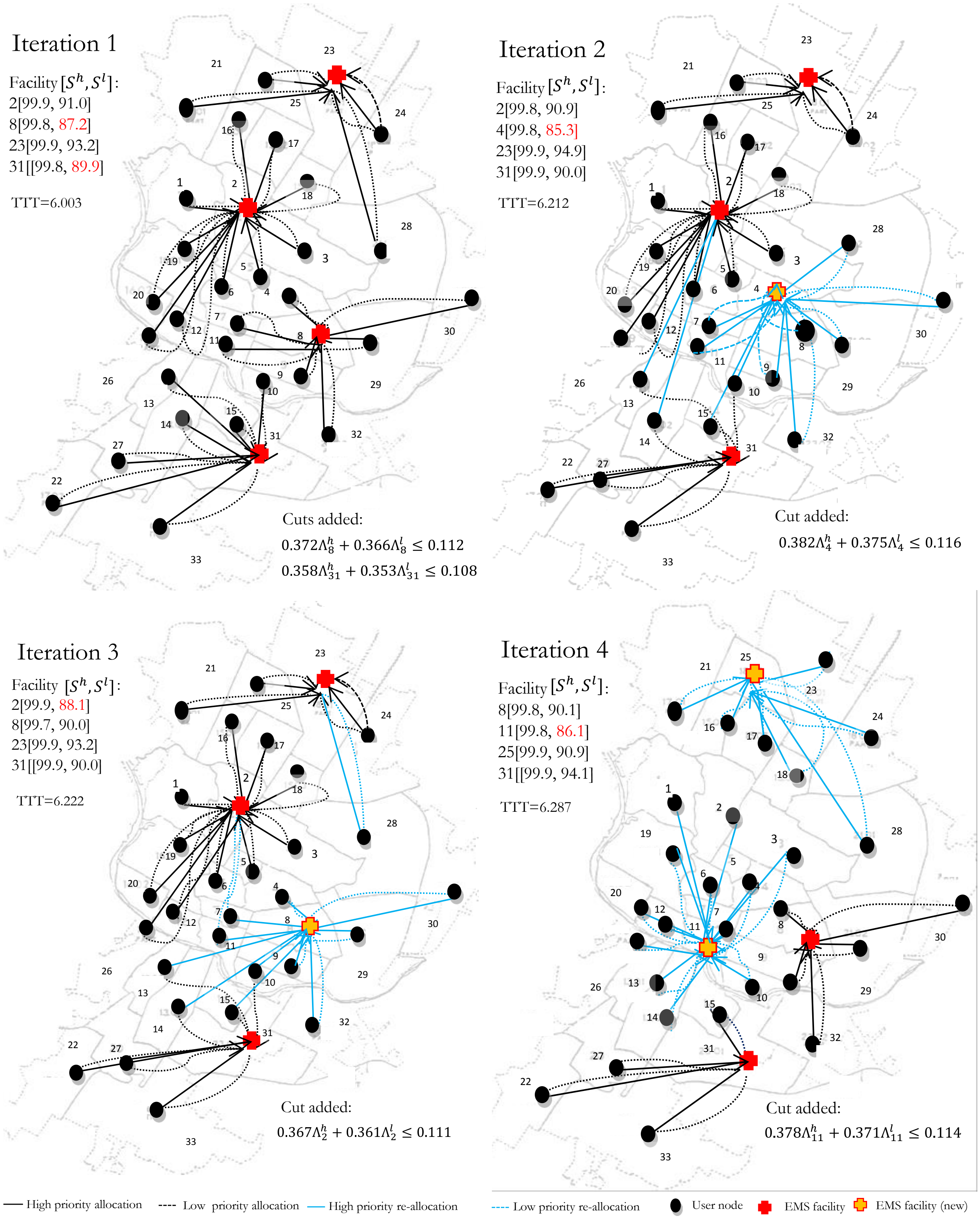

Figure 4: Illustrative Example: Iterations 1-4

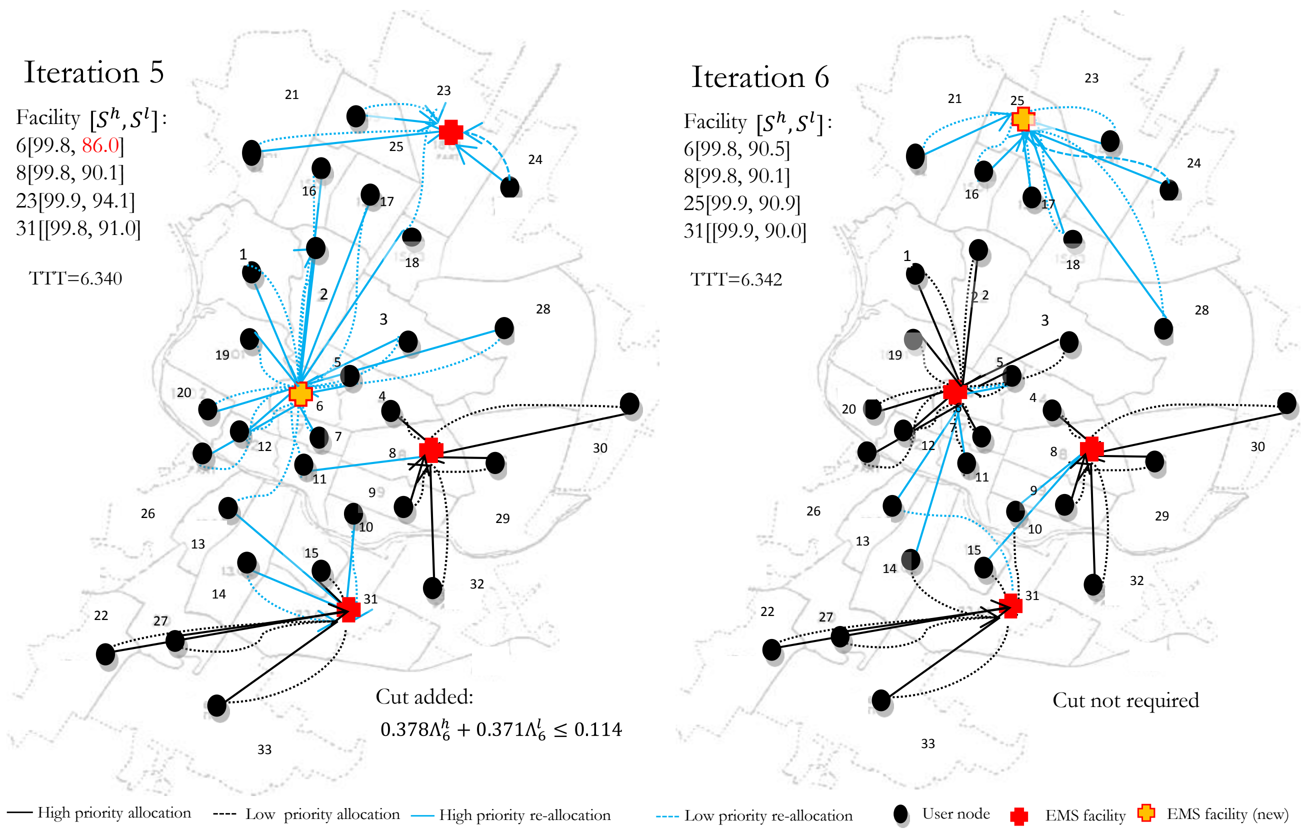

Figure 5: Illustrative Example: Iteration 5

Table 2: EMS Facility Location and Service Levels for the Model with Preemptive Priority under UC

f^h	f^l	τ^l	α^h	α^l	CPU	Iter.	NF	TT	Facility$[S^h, S^l]$								
0.5	99.5	15	98	80	3.2	1	4	6.003	2[99.9, 89.2]	8[99.9, 87.2]	23[99.9, 94.9]	31[99.9, 89.9]					
				85	3.3	1	4	6.003	2[99.9, 89.2]	8[99.9, 87.2]	23[99.9, 94.9]	31[99.9, 89.9]					
				90	92.1	12	5	5.512	3[99.9, 92]	8[99.9, 90.1]	11[99.9, 90.9]	23[100, 94.9]	31[100, 94.1]				
				95	241.5	16	9	4.254	2[100, 95.1]	8[100, 96.1]	9[100, 95.7]	11[100, 95.7]	14[100, 95.3]	17[100, 95.3]	23[100, 95.9]	29[100, 98]	31[100, 95.9]
		30		80	2.8	1	4	6.003	2[99.9, 92.9]	8[99.9, 91.5]	23[99.9, 96.8]	31[99.9, 93.4]					
				85	2.7	1	4	6.003	2[99.9, 92.9]	8[99.9, 91.5]	23[99.9, 96.8]	31[99.9, 93.4]					
				90	2.7	1	4	6.003	2[99.9, 92.9]	8[99.9, 91.5]	23[99.9, 96.8]	31[99.9, 93.4]					
				95	207.2	14	6	5.262	2[99.9, 95.4]	8[100, 97.3]	9[100, 96.1]	11[99.9, 95.3]	23[100, 95.7]	31[100, 96.2]			
		60		80	2.7	1	4	6.003	2[99.9, 96.9]	8[99.9, 96.2]	23[99.9, 98.7]	31[99.9, 97.2]					
				85	2.7	1	4	6.003	2[99.9, 96.9]	8[99.9, 96.2]	23[99.9, 98.7]	31[99.9, 97.2]					
				90	2.7	1	4	6.003	2[99.9, 96.9]	8[99.9, 96.2]	23[99.9, 98.7]	31[99.9, 97.2]					
				95	2.7	1	4	6.003	2[99.9, 96.9]	8[99.9, 96.2]	23[99.9, 98.7]	31[99.9, 97.2]					
		120		80	2.8	1	4	6.003	2[99.9, 99.4]	8[99.9, 99.2]	23[99.9, 99.8]	31[99.9, 99.5]					
				85	2.7	1	4	6.003	2[99.9, 99.4]	8[99.9, 99.2]	23[99.9, 99.8]	31[99.9, 99.5]					
				90	2.7	1	4	6.003	2[99.9, 99.4]	8[99.9, 99.2]	23[99.9, 99.8]	31[99.9, 99.5]					
				95	2.7	1	4	6.003	2[99.9, 99.4]	8[99.9, 99.2]	23[99.9, 99.8]	31[99.9, 99.5]					
1	99	15		80	2.6	1	4	6.003	2[99.9, 91]	8[99.8, 85.4]	23[99.9, 94.9]	31[99.8, 89.9]					
				85	2.6	1	4	6.003	2[99.9, 91]	8[99.8, 85.4]	23[99.9, 94.9]	31[99.8, 89.9]					
				90	79.0	12	5	5.512	3[99.9, 92]	8[99.8, 90.1]	11[99.9, 90.9]	23[99.9, 94.9]	31[99.9, 94.1]				
				95	211.3	16	9	4.254	2[99.9, 95.1]	8[99.9, 96.1]	9[99.9, 95.7]	11[99.9, 95.7]	14[100, 95.3]	17[99.9, 95.3]	23[99.9, 95.9]	29[100, 98]	31[99.9, 95.9]
		30		80	2.6	1	4	6.003	2[99.9, 94.1]	8[99.8, 90.1]	23[99.9, 96.8]	31[99.8, 93.4]					
				85	2.5	1	4	6.003	2[99.9, 94.1]	8[99.8, 90.1]	23[99.9, 96.8]	31[99.8, 93.4]					
				90	2.5	1	4	6.003	2[99.9, 94.1]	8[99.8, 90.1]	23[99.9, 96.8]	31[99.8, 93.4]					
				95	211.4	14	6	5.262	2[99.9, 95.4]	8[99.9, 97.3]	9[99.9, 96.1]	11[99.9, 95.3]	23[99.9, 95.7]	31[99.9, 96.2]			
		60		80	2.6	1	4	6.003	2[99.9, 97.5]	8[99.8, 95.5]	23[99.9, 98.7]	31[99.8, 97.2]					
				85	2.6	1	4	6.003	2[99.9, 97.5]	8[99.8, 95.5]	23[99.9, 98.7]	31[99.8, 97.2]					
				90	2.6	1	4	6.003	2[99.9, 97.5]	8[99.8, 95.5]	23[99.9, 98.7]	31[99.8, 97.2]					
				95	2.5	1	4	6.003	2[99.9, 97.5]	8[99.8, 95.5]	23[99.9, 98.7]	31[99.8, 97.2]					
		120		80	2.7	1	4	6.003	2[99.9, 99.6]	8[99.8, 99]	23[99.9, 99.8]	31[99.8, 99.5]					
				85	2.6	1	4	6.003	2[99.9, 99.6]	8[99.8, 99]	23[99.9, 99.8]	31[99.8, 99.5]					
				90	2.6	1	4	6.003	2[99.9, 99.6]	8[99.8, 99]	23[99.9, 99.8]	31[99.8, 99.5]					
				95	2.6	1	4	6.003	2[99.9, 99.6]	8[99.8, 99]	23[99.9, 99.8]	31[99.8, 99.5]					
5	95	15		80	2.4	1	4	6.003	2[99.2, 90.9]	8[99, 85.4]	23[99.6, 94.9]	31[99.2, 89.9]					
				85	2.4	1	4	6.003	2[99.2, 90.9]	8[99, 85.4]	23[99.6, 94.9]	31[99.2, 89.9]					
				90	19.8	5	4	6.309	6[99.2, 90.4]	8[99.2, 90]	25[99.3, 90.9]	31[99.4, 90]					
				95	201.7	15	8	5.852	3[99.4, 95.2]	5[99.8, 95]	12[99.5, 95.1]	14[99.6, 95.3]	23[99.7, 95.8]	28[99.7, 96.4]	31[99.6, 95]	32[99.8, 95]	
		30		80	2.4	1	4	6.003	2[99.2, 94]	8[99, 90.1]	23[99.6, 96.8]	31[99.2, 93.4]					
				85	2.4	1	4	6.003	2[99.2, 94]	8[99, 90.1]	23[99.6, 96.8]	31[99.2, 93.4]					
				90	2.4	1	4	6.003	2[99.2, 94]	8[99, 90.1]	23[99.6, 96.8]	31[99.2, 93.4]					
				95	184.9	14	6	5.262	2[99.4, 95.4]	8[99.7, 96.2]	9[99.5, 97.2]	11[99.4, 95.2]	23[99.5, 95.6]	31[99.5, 96.2]			
		60		80	2.7	1	4	6.003	2[99.2, 97.5]	8[99, 95.5]	23[99.6, 98.7]	31[99.2, 97.1]					
				85	2.5	1	4	6.003	2[99.2, 97.5]	8[99, 95.5]	23[99.6, 98.7]	31[99.2, 97.1]					
				90	2.5	1	4	6.003	2[99.2, 97.5]	8[99, 95.5]	23[99.6, 98.7]	31[99.2, 97.1]					
				95	2.4	1	4	6.003	2[99.2, 97.5]	8[99, 95.5]	23[99.6, 98.7]	31[99.2, 97.1]					
		120		80	2.5	1	4	6.003	2[99.2, 99.5]	8[99, 99]	23[99.6, 99.8]	31[99.2, 99.5]					
				85	2.5	1	4	6.003	2[99.2, 99.5]	8[99, 99]	23[99.6, 99.8]	31[99.2, 99.5]					
				90	2.5	1	4	6.003	2[99.2, 99.5]	8[99, 99]	23[99.6, 99.8]	31[99.2, 99.5]					
				95	2.7	1	4	6.003	2[99.2, 99.5]	8[99, 99]	23[99.6, 99.8]	31[99.2, 99.5]					

Table 3: Total Patient Travel Times in DC versus UC

								Directed Choice		User Choice		% Change
R	μ^h	μ^l	τ^l	f^h	f^l	α^h	α^l	NF	TT	NF	TT	TT
Preemptive Priority												
10	2	2	30	0.5	99.5	98	90	4	6.003	4	6.003	0.00
							95	5	5.815	6	5.262	9.50
				5	95		90	4	6.003	4	6.003	0.00
							95	5	5.763	6	5.262	8.70
		3		0.5	99.5		90	4	6.003	4	6.003	0.00
							95	4	6.003	4	6.003	0.00
				5	95		90	4	6.003	4	6.003	0.00
							95	4	6.003	4	6.003	0.00
	3	2		0.5	99.5		90	4	6.003	4	6.003	0.00
							95	5	5.809	6	5.262	9.41
				5	95		90	4	6.003	4	6.003	0.00
							95	5	5.695	5	5.976	-4.93
		3		0.5	99.5		90	4	6.003	4	6.003	0.00
							95	4	6.003	4	6.003	0.00
				5	95		90	4	6.003	4	6.003	0.00
							95	4	6.003	4	6.003	0.00
Non-preemptive Priority												
10	12	12	15	0.5	99.5	98	90	5	5.855	6	5.262	10.13
							95	5	5.855	6	5.262	10.13
				5	95		90	5	5.800	6	5.262	9.28
							95	5	5.800	6	5.262	9.28
		15		0.5	99.5		90	4	6.783	5	5.698	16.00
							95	4	6.783	5	5.698	16.00
				5	95		90	4	6.719	5	5.698	15.19
							95	4	6.719	5	5.698	15.19
	15	12		0.5	99.5		90	5	5.855	6	5.262	10.12
							95	5	5.855	6	5.262	10.12
				5	95		90	5	5.784	6	5.262	9.02
							95	5	5.784	6	5.262	9.02
		15		0.5	99.5		90	4	6.783	5	5.698	15.99
							95	4	6.783	5	5.698	15.99
				5	95		90	4	6.359	5	5.698	10.39
							95	4	6.359	5	5.698	10.39

6 Conclusion and Future Research

We studied the problem of locating EMS facilities, which are 24 hour, 7 days-a-week, medical facilities focussed on the delivery of ambulatory care to treat injuries or illnesses requiring immediate care. Motivated by the development of acuity rating systems (like ESI in US, CTAS in Canada, ATS in Australia), which are meant to help EMS service providers correctly triage patients into different acuity classes, we studied the problem in the presence of heterogeneous patients, belonging to different classes. To the best of our knowledge, ours is the first study on a location-allocation problem in presence of heterogeneous customers with a different service level requirement for each class, and where the service level constraint for each customer class is defined using the complete distribution of its waiting time, as opposed to its average waiting time at an EMS facility. We modeled the network of EMS facilities as spatially distributed M/M/1 priority queues, whose locations and user allocations need to be determined. The resulting integer programming problem was challenging to solve, especially in absence of any known analytical expression for the waiting time distribution of low priority customers in an M/M/1 priority queue. We developed a cutting plane based solution algorithm, exploiting the concavity of the waiting time distribution of low priority customers to approximate its non-linearity using tangent planes, determined numerically using matrix geometric method.

CTAS, ESI and ATS use 5-level triage acuity scales. However, in the current paper, we assumed only two priority classes (high and low priority) for emergency patients, primarily for tractability. We see extension of the current work to more than 2 priority classes of emergency patients as a possible, yet challenging, direction for future research. The current work can also be extended for general, instead of exponential, service time distribution at EMS facilities.

ACKNOWLEDGMENT

This research was supported by the Research & Publication Grant, Indian Institute of Management Ahmedabad, provided to the first author, and by the National Science and Engineering Research Council of Canada (NSERC) grant, provided to the second author.

References

References

Aboolian, R., Berman, O., and Krass, D. (2012). Profit maximizing distributed service system design with congestion and elastic demand. *Transportation Science*, 46(2):247–261.

Amiri, A. (1997). Solution procedures for the service system design problem. *Computers & operations research*, 24(1):49–60.

Baron, O., Berman, O., and Krass, D. (2008). Facility location with stochastic demand and constraints on waiting time. *Manufacturing & Service Operations Management*, 10(3):484–505.

Belotti, P., Labbé, M., Maffioli, F., and Ndiaye, M. M. (2007). A branch-and-cut method for the obnoxious p-median problem. *4OR*, 5(4):299–314.

Berman, O. and Krass, D. (2002). Facility location problems with stochastic demands and congestion. In Drezner, Z. and Hamacher, H., editors, *Facility Location: Applications and Theory*. Springer.

Berman, O., Krass, D., and Wang, J. (2006). Locating service facilities to reduce lost demand. *IIE Transactions*, 38(11):933–946.

Boffey, B., Galvao, R., and Espejo, L. (2007). A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*, 178(3):643–662.

Cánovas, L., García, S., Labbé, M., and Marín, A. (2007). A strengthened formulation for the simple plant location problem with order. *Operations Research Letters*, 35(2):141–150.

Castillo, I., Ingolfsson, A., and Sim, T. (2009). Socially optimal location of facilities with fixed servers, stochastic demand and congestion. *Production and Operations Management*, 18(6):721–736.

Church, R. L. and Cohon, J. L. (1976). Multiobjective location analysis of regional energy facility siting problems. Technical report, Brookhaven National Lab., Upton, NY (USA).

Daskin, M. S. (1982). Application of an expected covering model to emergency medical service system design. *Decision Sciences*, 13(3):416–439.

Daskin, M. S. and Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152.

Dobson, G. and Karmarkar, U. S. (1987). Competitive location on a network. *Operations Research*, 35(4):565–574.

Elhedhli, S. (2006). Service system design with immobile servers, stochastic demand, and congestion. *Manufacturing & Service Operations Management*, 8(1):92–97.

Espejo, I., Marín, A., and Rodríguez-Chía, A. M. (2012). Closest assignment constraints in discrete location problems. *European Journal of Operational Research*, 219(1):49–58.

Gilboy, N., Tanabe, T., Travers, D., and Rosenau, A. (2011). Emergency Severity Index (ESI): A triage tool for emergency department care, version 4. Implementation Handbook 2012 Edition.

Jayaswal, S., Jewkes, E., and Ray, S. (2011). Product differentiation and operations strategy in a capacitated environment. *European Journal of Operational Research*, 210(3):716–728.

Latouche, G. and Ramaswai, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM Series on Statistics and Applied Probability.

Marianov, V. and Serra, D. (1998). Probabilistic, maximal covering locationallocation models forcongested systems. *Journal of Regional Science*, 38(3):401–424.

Marianov, V. and Serra, D. (2002). Location–allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research*, 111(1-4):35–50.

Marín, A. (2011). The discrete facility location problem with balanced allocation of customers. *European Journal of Operational Research*, 210(1):27–38.

Murray, J. M. (2003). The canadian triage and acuity scale: A canadian perspective on emergency department triage. *Emergency Medicine*, 15(1):6–10.

Nair, R. and Miller-Hooks, E. (2009). Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, 2137(1):63–73.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Dover Publications.

Ramaswami, V. and Lucantoni, D. M. (1985). Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes. *Communications in Statistics. Stochastic Models*, 1(2):125–136.

Rojeski, P. and ReVelle, C. (1970). Central facilities location under an investment constraint. *Geographical Analysis*, 2(4):343–360.

Silva, F. and Serra, D. (2008). Locating emergency services with different priorities: the priority queuing covering location problem. *Journal of the Operational Research Society*, 59(9):1229–1238.

Vidyarthi, N. and Jayaswal, S. (2014). Efficient solution of a class of location–allocation problems with stochastic demand and congestion. *Computers & Operations Research*, 48:20–30.

Vidyarthi, N. and Kuzgunkaya, O. (2014). The impact of directed choice on the design of preventive healthcare facility network under congestion. *Health Care Management Science*, pages DOI 10.1007/s10729–014–9274–2.

Wagner, J. and Falkson, L. (1975). The optimal nodal location of public facilities with price-sensitive demand. *Geographical Analysis*, 7(1):69–83.

Wang, Q., Batta, R., and Rump, C. M. (2002). Algorithms for a facility location problem with stochastic customer demand and immobile servers. *Annals of Operations Research*, 111(1-4):17–34.

Zhang, Y., Berman, O., and Verter, V. (2012). The impact of client choice on preventive healthcare facility network design. *OR spectrum*, 34(2):349–370.

Appendix

1. Infinitesimal generator sub-matrices under non-preemptive priority

$$
L_0 = \left(\begin{array}{c|ccccc}
 & (0,0) & (0,1) & (0,2) & (0,\ldots) & (0,M) \\ \hline
(0,0) & * & \Lambda_j^h & & & \\
(0,1) & \mu_j^h & * & \Lambda_j^h & & \\
(0,2) & & \mu_j^h & * & \Lambda_j^h & \\
(0,\ldots) & & & \ddots & \ddots & \ddots \\
(0,M) & & & & \mu_j^h & *
\end{array} \right)
$$

$$
F_0 = \left(\begin{array}{c|ccccccccc}
 & (1,0) & (1,h,1) & (1,h,2) & (1,h,\ldots) & (1,h,M) & (1,l,1) & (1,l,2) & (1,l,\ldots) & (1,l,M) \\ \hline
(0,0) & \Lambda_j^l & & & & & & & & \\
(0,1) & & \Lambda_j^l & & & & & & & \\
(0,2) & & & \Lambda_j^l & & & & & & \\
(0,\ldots) & & & & \ddots & & & & & \\
(0,M) & & & & & \Lambda_j^l & & & &
\end{array} \right)
$$

$$
B_0 = \left(\begin{array}{c|ccccc}
 & (0,0) & (0,1) & (0,2) & (0,\ldots) & (0,M) \\ \hline
(1,0) & \mu_j^l & & & & \\
(1,h,1) & & & & & \\
(1,h,2) & & & & & \\
(1,h,\ldots) & & & & & \\
(1,h,M) & & & & & \\
(1,l,1) & 0 & \mu_j^l & & & \\
1,l,2) & & & \mu_j^l & & \\
(1,l,\ldots) & & & & \ddots & \\
(1,l,M) & & & & & \mu_j^l
\end{array} \right)
$$

$$
L = \left(\begin{array}{c|ccccccccc}
 & (k,0) & (k,h,1) & (k,h,2) & (k,h,\ldots) & (k,h,M) & (k,l,1) & (k,l,2) & (k,l,\ldots) & (k,l,M) \\ \hline
(k,0) & * & \Lambda_j^h & & & & & & & \\
(k,h,1) & \mu_j^h & * & \Lambda_j^h & & & & & & \\
(k,h,2) & & \mu_j^h & * & \Lambda_j^h & & & & & \\
(k,h,\ldots) & & & \ddots & \ddots & \ddots & & & & \\
(k,h,M) & & & & \mu_j^h & * & 0 & & & \\
(k,l,1) & & & & & 0 & * & \Lambda_j^l & & \\
(k,l,2) & & & & & & 0 & * & \Lambda_j^l & \\
(k,l,\ldots) & & & & & & & \ddots & \ddots & \ddots \\
(k,l,M) & & & & & & & & 0 & *
\end{array} \right)
$$

$$
F = \left(\begin{array}{c|ccccccccc}
 & (k+1,0) & (k+1,h,1) & (k+1,h,2) & (k+1,h,\ldots) & (k+1,h,M) & (k+1,l,1) & (k+1,l,2) & (k+1,l,\ldots) & (k+1,l,M) \\ \hline
(k,0) & \Lambda_j^l & & & & & & & & \\
(k,h,1) & & \Lambda_j^l & & & & & & & \\
(k,h,2) & & & \Lambda_j^l & & & & & & \\
(k,h,\ldots) & & & & \ddots & & & & & \\
(k,h,M) & & & & & \Lambda_j^l & & & & \\
(k,l,1) & & & & & & \Lambda_j^l & & & \\
(k,l,2) & & & & & & & \Lambda_j^l & & \\
(0,l,\ldots) & & & & & & & & \ddots & \\
(0,l,M) & & & & & & & & & \Lambda_j^l
\end{array} \right)
$$

$$B = \left(\begin{array}{c|ccccccccc}
 & (k,0) & (k,h,1) & (k,h,2) & (k,h,\ldots) & (k,h,M) & (k,l,1) & (k,l,2) & (k,l,\ldots) & (k,l,M) \\
\hline
(k-1,0) & \mu_j^l & & & & & & & & \\
(k-1,h,1) & & & & & & & & & \\
(k-1,h,2) & & & & & & & & & \\
(k-1,h,\ldots) & & & & & & & & & \\
(k-1,h,M) & & & & & & & & & \\
(k-1,l,1) & 0 & \mu_j^l & & & & & & & \\
(k-1,l,2) & & & \mu_j^l & & & & & & \\
(k-1,l,\ldots) & & & & \ddots & & & & & \\
(k-1,l,M) & & & & & \mu_j^l & & & &
\end{array}\right)$$

where $*$ is such that $A_0\mathbf{e} + B_0\mathbf{e} = \mathbf{0}$. $A_1 = B_0 - A_2$.

2. Data

Table A1: 5-Month Period Census Tract Level Service Call Data for Austin, Texas (Daskin and Stern, 1981)

Zone Number	Census Tract	EMS calls in a 5-month period	Zone Number	Census Tract	EMS calls in a 5-month period	Zone Number	Census Tract	EMS calls in a 5-month period
1	1	72	12	12	48	23	18.01	246
2	2	176	13	13.01	105	24	18.02	102
3	3	193	14	13.02	232	25	18.03	120
4	4	137	15	14	133	26	19	36
5	5	32	16	15.01	56	27	20	202
6	6	96	17	15.02	104	28	21.01	182
7	7	83	18	15.03	81	29	21.02	190
8	8	317	19	16.01	86	30	22	46
9	9	299	20	16.02	20	31	23.01	128
10	10	98	21	17.01	115	32	23.02	100
11	11	207	22	17.02	59	33	24	148

Table A2: Census Tract Level Travel Time (in minutes) Data for Austin, Texas (Daskin and Stern, 1981)

From ↓ \ To →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	1	5	8	11	8	8	9	12	13	13	9	8	11	12	13	6	9	10	6	7	9	16	14	16	11	10	15	13	16	18	15	19	22
2		1	4	7	4	4	6	9	11	10	7	7	9	10	11	7	7	7	5	6	11	15	12	13	10	9	14	10	12	14	13	16	19
3			1	6	5	5	6	7	9	9	7	11	12	12	10	8	8	6	9	10	13	19	11	10	10	13	16	7	10	12	12	14	18
4				1	6	5	4	3	5	7	5	9	10	10	7	13	12	9	10	9	17	17	13	12	13	12	14	9	7	9	10	10	16
5					1	3	5	7	9	9	6	9	10	10	9	9	10	8	7	8	13	17	14	13	12	11	15	9	11	13	11	14	18
6						1	3	6	8	7	4	6	8	8	8	10	10	9	6	6	14	16	14	13	13	9	13	10	11	13	10	14	16
7							1	4	6	5	2	5	7	7	6	11	12	10	7	6	14	15	14	13	14	9	11	10	9	12	8	12	14
8								1	4	5	4	8	9	9	7	14	13	11	10	9	18	17	14	14	14	12	13	10	7	10	9	9	15
9									1	4	5	8	9	9	6	15	15	13	11	10	19	16	17	16	16	12	13	12	7	11	8	8	13
10										1	4	7	7	6	3	15	15	12	11	9	18	13	16	15	16	11	10	12	10	14	5	10	12
11											1	4	5	5	5	11	13	11	7	5	15	13	15	14	15	7	9	11	9	13	7	11	13
12												1	5	7	7	10	13	13	6	4	14	13	18	17	15	6	11	14	13	16	10	14	16
13													1	3	6	13	15	14	9	7	16	10	19	18	17	7	8	15	14	17	8	14	14
14														1	5	14	16	15	10	8	17	10	19	18	19	8	7	15	14	17	6	13	12
15															1	15	16	13	10	9	18	12	17	16	17	10	9	13	11	15	4	10	11
16																1	3	6	8	9	7	18	11	12	6	12	17	10	17	16	18	21	24
17																	1	5	10	11	9	21	10	11	6	14	19	9	16	15	18	20	24
18																		1	11	12	12	21	9	10	7	15	20	6	14	12	15	17	21
19																			1	4	11	14	16	17	12	7	13	14	15	17	13	17	19
20																					12	12	17	18	13	6	11	15	14	17	11	16	18
21																					1	22	10	14	7	15	20	15	21	20	21	25	27
22																						1	26	26	23	12	7	23	21	25	10	19	12
23																							1	6	6	20	24	10	17	14	19	21	25
24																								1	10	21	23	6	14	9	18	17	24
25																									1	16	21	10	17	15	19	21	25
26																										1	11	18	17	20	12	17	19
27																											1	19	18	22	7	16	12
28																												1	11	8	15	15	21
29																													1	9	13	9	16
30																														1	17	11	19
31																															1	11	8
32																																1	13
33																																	1