European Journal of Operational Research 309 (2023) 634–655

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

Multiple allocation hub location with service level constraints for two shipment classes

Sachin Jayaswal[a], Navneet Vidyarthi[b,*]

[a] Indian Institute of Management, Vastrapur, Ahmedabad, Gujarat 380 015, India
[b] Department of Supply Chain and Business Technology Management, John Molson School of Business and CIRRELT, Concordia University, Montreal, QC H3E 2A3, Canada

ARTICLE INFO

Article history:
Received 27 December 2021
Accepted 31 January 2023
Available online 7 February 2023

Keywords:
Location
Hub networks
Service level
Priority queue
Cutting plane method

ABSTRACT

In this paper, we study a hub network design problem arising in the context of a third-party logistics (3PL) service provider, which acts as an intermediary between shippers and carriers. A 3PL service provider usually caters to different classes of shipments that require different levels of service, e.g. two-day delivery, next-day delivery etc. We, therefore, study the problem under stochastic demand from two classes of shipments, with one class receiving priority over the other in service at the hubs to maintain the different service levels required by them. To this end, we present two models for designing a capacitated hub network with a service level constraint, defined using the distribution of time spent at hubs, for each shipment class. The models seek to design the hub network at the minimum total cost, which includes the total fixed cost of equipping open hubs with sufficient processing capacity and the variable transportation costs. The network of hubs, given their locations, is thus modeled as spatially distributed priority queues. The resulting model is challenging to solve, for which we propose a cutting plane-based exact solution method.

1. Introduction

We consider a hub network design problem arising in the context of a third-party logistics (3PL) service provider, which acts as an intermediary between shippers and carriers. A 3PL service provider realizes economies of scale by first shipping generally less-than-truckload (LTL) shipments from a shipper to a hub (consolidation center) where they are consolidated with generally LTL shipments from several other shippers into a full truckload (TL). The consolidated TL is then shipped to a second hub (deconsolidation center) before deconsolidating and distributing its components to their final destinations. Due to economies of scale, the consolidated TL shipments between two hubs (consolidation and deconsolidation centers) are relatively less expensive per unit distance compared to LTL shipments between the origins/destinations and the consolidation/deconsolidation centers. However, routing a shipment through intermediate hubs instead of shipping it directly from its origin to its destination generally increases the distance travelled, and may possibly increase the total shipment cost in the absence of any discount arising from consolidation of shipments. Moreover, opening consolidation/deconsolidation centers entails cost of equipping them with human resources and other resources for unloading, sorting, loading, and temporary storage (Uster & Agrahari, 2011). Hence, the hub network design calls for finding the optimal number and locations of hubs, and the routes for the shipments from their origins to the destinations through hubs such that the total shipment cost and the hub location cost is minimized.

The hub network design problem for a 3PL service provider, as described above, can be classified as a variant of *hub location problems with fixed costs*. Refer to Alumur & Kara (2008); Alumur et al. (2021); Campbell et al. (2002); Campbell & O'Kelly (2012); Contreras & O'Kelly (2019), and Contreras (2021) for comprehensive reviews of the related literature. O'Kelly (1992b) introduced the hub location problem with fixed costs, in which the number of hubs to open is a decision variable as opposed to a *p*-hub median problem in which the number of hubs to open is given. His model assumes single allocation in the sense that all the shipments originating at a given node are always routed through the same first hub (consolidation center), irrespective of their destinations. Similarly, all the shipments destined to a given node are always received from the same second hub (deconsolidation center), irrespective of their origins. Following O'Kelly's work, several

* Corresponding author.
E-mail addresses: sachin@iima.ac.in (S. Jayaswal), n.vidyarthi@concordia.ca (N. Vidyarthi).

https://doi.org/10.1016/j.ejor.2023.01.066
0377-2217/

papers have reported either different formulations or different solution approaches to the single allocation hub location problem with fixed costs (Abdinnour-Helm & Venkataramanan, 1998; Alumur et al., 2009; 2012; Campbell, 1994b; Contreras et al., 2012; Cunha & Silva, 2007; Labbe & Yaman, 2004). Simultaneously, there are papers that have studied the multiple allocation versions of the problem, in which the shipments originating at a given node may be routed through different first hubs, depending on their destinations (likewise the shipments destined to a given node may be received from different second hubs, depending on their origins) (Alumur et al., 2012; Boland et al., 2004; Campbell, 1994b; Canovas et al., 2007; Hamacher et al., 2004; Marin, 2005b; Marıń et al., 2006; Racunica & Wynter, 2005).

In reality, the hubs have a finite limit on the amount of traffic they can handle. A consolidation center, for example, may have a limit on the amount of shipments it can unload, sort or load in a given time period before distributing them. Thus, the solution of a cost minimization hub location problem, when implemented, may cause traffic delays due to congestion. Explicit modeling of such a limit on the amount of flows that a hub can handle leads to a capacitated hub location problem. Aykin (1994); Costa et al. (2008); Ernst & Krishnamoorthy (1999); Labbe et al. (2005), and Bhatt et al. (2021); Correia et al. (2010), among others, have studied the single allocation version, while Boland et al. (2004); Ebery et al. (2000), and Marin (2005a) have dealt with the multiple allocation version of the capaciatated hub location problem.

Although a 3PL service provider may wish to follow scheduled pickups of shipments from different shippers, they are often subjected to delays at the pickup locations/ consolidation centers, making their arrivals at the consolidation/ deconsolidation centers non-deterministic. The service times of the shipments at the hubs also exhibit variability. For example, unloading a truck with only a few (lighter) shipments at a hub may take much less time than a truck carrying too many (bulky) shipments. Further, service requirements of different shipments arriving at a hub may themselves be not homogeneous. While the service process of a shipment at its first hub (consolidation center) generally consists of (i) unloading, (ii) batching, and (iii) loading, it does not need batching if its origin and destination are both served by the same hub. On the other hand, a hub that is a consolidation center for one shipment may be a deconsolidation center (second hub) for another shipment, in which case its service process at that hub will consist of (i) unloading, (ii) bulk-breaking, and (iii) loading (Ishfaq & Sox, 2012). This difference in the service process for different shipments at hubs creates variability in their service times, making them non-deterministic. The service times of shipments may also vary due to variability in the operating conditions at hubs (Marianov & Serra, 2003).

Variability in the arrival and service processes of shipments creates congestion at hubs, resulting in delays. O'Kelly (1986a) emphasized possible negative repercussions of highly utilized hubs and suggested the minimization of the variability of hub usage. Congestion at hubs has also been addressed by Alkaabneh et al. (2019); Azizi et al. (2018); Bhatt et al. (2021); Camargo et al. (2009, 2011b); Elhedhli & Hu (2005); Elhedhli & Wu (2010); Guldmann & Shen (1997); Hasanzadeh et al. (2018); Kian & Kargar (2016); Marianov & Serra (2003); Mohammadi et al. (2011), among others. Guldmann & Shen (1997) is one of the early papers to model hubs as M/M/1 queuing systems. Marianov & Serra (2003) and Mohammadi et al. (2011) modeled hubs as M/D/c and M/M/c queuing systems, respectively, and used probabilistic service level constraints to limit congestion (expressed as the number of shipments (airplanes) waiting) at hubs. Others addressed congestion at hubs by imposing an increasing penalty for each incremental unit of traffic flow at a hub. For this, Alkaabneh et al. (2019); Camargo et al. (2009); Elhedhli & Hu (2005); Kian & Kargar (2016) used a convex power-law congestion penalty, whereas Azizi et al. (2018); Bhatt et al. (2021); Camargo et al. (2011a); Elhedhli & Wu (2010); Hasanzadeh et al. (2018) used queuing-based congestion penalty functions. Table 1 provides a summary of the literature on the hub location problem with congestion.

To the best of our knowledge, Marianov & Serra (2003) and Mohammadi et al. (2011) are the only two papers to have accounted for congestion at hubs by explicitly imposing probabilistic service level constraints (defined in terms of probability of waiting) at hubs. They assume the same treatment of all shipments at hubs. A slightly related work is hub location for time-definite transportation by Campbell (2009), which imposes a constraint on the maximum time a shipment can take from its origin to its destination. However, it does not account for the delay due to service activities such as unloading, sorting, batching, and loading at hubs. Neither does it account for the delay caused by congestion, arising from the limited capacity at hubs, nor does it, like Marianov & Serra (2003) and Mohammadi et al. (2011), account for the shipments with different service level requirements. However, in reality, shipments arriving at a hub are not homogeneous: some of them may have longer delivery time windows, for example, two-day delivery (henceforth called 'regular') shipments, while others may have a much shorter delivery time window, for example, the next-day delivery (henceforth called 'express') shipments. For example, flipkart.com, one of the most popular e-retailers in India, offers the express same-day and next-day delivery options in select cities, besides the standard delivery option.[1] Similarly, Amazon offers the express one-day delivery and two-day delivery options for selected products in select cities, besides the standard delivery option. These two classes of shipments, with different delivery time guarantees, when handled by a 3PL service provider, will require different treatments at the hubs, with the express shipments deserving a priority over the regular shipments. Our study takes such heterogeneous shipments into account by imposing a different service level requirement for each shipment class at any hub, with a more stringent service level, and hence a priority service, for the express class. We define the service level requirement for a shipment class in terms of the minimum probability with which its dwell time at a hub should be within a pre-defined threshold.

To the best of our knowledge, this is the first paper on the design of hub networks to consider the trade-offs between costs and service for heterogeneous shipment classes requiring different service levels in the presence of congestion arising at hubs. We present a hub node location model and a hub arc location model where hubs are modeled as preemptive priority M/M/1 queues. The resulting mixed integer programming (MIP) problems with probabilistic constraints are challenging to solve, especially in the absence of any known closed-form expression for the service level constraint for low-priority customers (regular shipments). To resolve this problem, we compute the service level for low-priority customers numerically using the matrix geometric method. We exploit the concavity of the sojourn time distribution of low-priority shipments to eliminate the non-linearity in their service level function, at the expense of a large number of tangent hyperplanes. The resulting model is solved efficiently using a cutting plane method, wherein we start with a set of linearization constraints and add the rest as needed. The use of the matrix geometric method to numerically compute the service level for low-priority customers at a hub is inspired by Jayaswal & Vidyarthi (2017), who used it to compute a slightly different service level at a service facility in a facility location model – their service level is based on the distribution of waiting time at a service facility, as opposed to that based on the distribution of the dwell (wait + service) time at a

[1] http://www.flipkart.com/faster-delivery

Table 1
Summary of the literature on the hub location problems with congestion.

Reference	Congestion model	Service level constraints	Demand/shipment class	Solution method
Guldmann & Shen (1997)	M/M/1 queue	–	Homogenous/One	Approximate (Piecewise Linearization)
Marianov & Serra (2003)	M/D/c queue	✓	Homogenous/One	Tabu search based heuristic
Elhedhli & Hu (2005)	Power-law function	–	Homogenous/One	Lagrangean heuristic
Camargo et al. (2009)	Power-law function	–	Homogenous/One	Exact (Benders decomposition)
Elhedhli & Wu (2010)	M/M/1 queue	–	Homogenous/One	Lagrangean heuristic
Camargo et al. (2011b)	M/M/1 queue, Power-law function	–	Homogenous/One	Exact (Outer approximation, Benders decomposition)
Mohammadi et al. (2011)	M/M/c queue	✓	Homogenous/One	Meta-heuristic
Kian & Kargar (2016)	Power-law function	–	Homogenous/One	Exact (Conic reformulations)
Azizi et al. (2018)	M/G/1 queue	–	Homogenous/One	Cutting plane method, GA
Alkaabneh et al. (2019)	Power-law function	–	Homogenous/One	Lagrangean heuristic, GRASP
Bhatt et al. (2021)	M/G/1 queue	–	Homogenous/One	Exact (Conic reformulation)
This paper	M/M/1 queue	✓	Heterogenous/Two	Exact (Cutting plane method, Matrix geometric method)

hub. The context of our problem, which is hub location, is also fundamentally different from an emergency facility location problem studied by Jayaswal & Vidyarthi (2017). The hub location problem studied in this paper is a difficult class of NP-hard discrete location problems that also involves routing decisions besides location. The issue of inter-hub flows, and hence the associated discount that arises in a hub location model is absent in a standard facility location model. As such, the insights from this paper do not apply to a facility location context.

1.1. Contribution

The contribution of this paper is three-fold. First, motivated by the growth of rapid e-commerce delivery services such as two-day delivery, next-day delivery, same-day delivery, etc., where different shipments require different processing and/or treatments at the hub facilities, this paper presents hub location models that account for operational considerations at hubs by incorporating probabilistic service level constraints for two classes of shipments, viz. regular and express. To the best of our knowledge, this is the first paper to account for the heterogeneity in shipments at hubs while designing the hub network. Second, hub location problems are a difficult class of NP-hard combinatorial optimization problems. The resulting hub location models with probabilistic service level constraints are challenging to solve, even for a single shipment class. Hence, the only two existing papers that have used such service level constraints in hub location models have resorted to heuristics (see Marianov & Serra, 2003; Mohammadi et al., 2011, in Table 1). With two shipment classes, the problem becomes even further challenging due to the absence of any known closed-form expression for the probabilistic service level constraint for low-priority customers (regular shipments). To this end, we propose an exact cutting plane-based method to solve the problem efficiently. Third, based on our extensive computational study of two classes of hub location models – hub node location model and hub arc location model – we present interesting managerial insights into the design of hub network in the presence of service level constraints for different shipment classes. Specifically, we show that extending priority service to a larger proportion of the customer base does not necessarily come at a cost. What is even more interesting is that, on the contrary, it may even reduce the cost.

The remainder of the paper is organized as follows. In Section 2, we present the models, followed by a discussion on the solution methodology in Section 3. Section 4 presents our computational study and discussion of results. The paper concludes with a summary of results and a discussion on future research in Section 5.

2. Model formulation

Let N be the set of nodes representing the origins and destinations of the shipments to be delivered from various shippers. Define λ_{ij} as the amount of traffic (number of shipments per unit time) to be routed from the origin $i \in N$ to the destination $j \in N$. To exploit the economies of scale, the 3PL service provider first ships LTL shipments from an origin node $i \in N$ to a hub $k \in N$ where they are consolidated with LTL shipments originating from several other nodes $i \in N$ into a TL, which is then shipped to a second hub $m \in N$. The transportation cost per unit of traffic from node i to node j routed via hubs k and m, in that order, is given by $C_{ijkm} = \delta C_{ik} + \alpha C_{km} + \gamma C_{mj}$, where δC_{ik} is the unit collection cost from the origin node i to the hub k; γC_{mj} is the unit distribution cost from the hub m to the destination node j; αC_{km} is the unit inter-hub transfer cost, and $\delta, \alpha, \gamma \in (0, 1)$ are the discount factors, reflecting economies of scale, on the collection links (spokes to hubs), inter-hub links (hubs to hubs), and the distribution links (hubs to spokes), respectively. Generally, $\delta < \alpha, \gamma < \alpha$ due to greater consolidation of shipments at hubs, leading to TL shipments on inter-hub links. Further, let F_k be the amortized cost of establishing a hub at node $k \in N$. The problem facing the 3PL service provider is to optimally decide the appropriate nodes $k, m \in N$ to locate hubs, and path(s) between all origin and destination pairs (i, j) such that every path traverses one or more hubs to benefit from the consolidation at hubs. To this end, let the binary variable $z_k = 1$ represent the location of a hub at node k; 0 otherwise. Let the variable $x_{ijkm} \geq 0$ represent the fraction of the total traffic from node i to node j routed via hubs located at nodes k and m, in that order. With these notations, we first present one of the strongest known formulations of the *Uncapacitated Multiple Allocation Hub Location Problem* (UMAHLP), proposed by Hamacher et al. (2004), since our proposed model builds on it:

$$[\text{UMAHLP}]: \min \quad \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{m\in N} C_{ijkm}\lambda_{ij}x_{ijkm} + \sum_{k\in N} F_k z_k \tag{1}$$

$$\text{s.t.} \quad \sum_{k\in N}\sum_{m\in N} x_{ijkm} = 1 \qquad \forall i, j \in N \tag{2}$$

$$\sum_{m\in N} x_{ijkm} + \sum_{m\in N\setminus\{k\}} x_{ijmk} \leq z_k \qquad \forall i, j, k \in N \tag{3}$$

$$x_{ijkm} \geq 0 \qquad \forall i, j, k, m \in N \tag{4}$$

$$z_k \in \{0, 1\} \qquad \forall k \in N \tag{5}$$

The objective function (1) minimizes the sum of the total transportation costs between all the origin-destination node pairs and the amortized cost of establishing all the hubs. Constraint set (2) requires that the traffic demand between any pair of nodes be completely satisfied. Constraint set (3) prohibits traffic from being routed via any intermediate node that is not a hub. Constraints (4) and (5) are non-negativity and integrality requirements.

Certain applications impose a further restriction of at most two hubs en route any path from an origin node to a destination node, including the origin or the destination if either of them itself is a hub. For example, postal services may require that any mail should not visit more than two post offices before its final destination (Marıń et al., 2006). Similarly, it may not be desirable for a passenger aircraft to stop at more than 2 hubs for long distance flights. Such a restriction is implicitly taken care of by the model (1)–(5) if: (a) the transportation costs satisfy the triangle inequality, i.e., $C_{ij} < C_{ik} + C_{kj}$, (b) there is no cost for setting the inter-hub links, and (c) the inter-hub discount factor (α) is the same between all pairs of hubs. This is so because under the above conditions, any shipment routed via three hubs k, l and m, in that order, is costlier (incurs additional transportation cost without any additional inter-hub flow discount) than the flow routed directly from hub k to hub m (Marıń et al., 2006). However, the transportation costs may not always satisfy the triangle inequality, especially if they are not proportional to distances. In such cases, the restriction of at most 2 hubs on any feasible path from an origin node to a destination node needs to be explicitly imposed through the following additional constraints (Camargo et al., 2009):

$$x_{ijij} \geq z_i + z_j - 1 \qquad \forall i, j \in N \tag{6}$$

$$\sum_{m \in N \setminus \{j\}} x_{ijim} \geq z_i - z_j \qquad \forall i, j \in N \tag{7}$$

$$\sum_{k \in N \setminus \{i\}} x_{ijkj} \geq z_j - z_i \qquad \forall i, j \in N \tag{8}$$

Constraint set (6) restricts any flow from the origin node i to the destination node j to travel only via the path $i \to i \to j \to j$ if both i and j are hubs. Constraint sets (7) and (8) ensure that any flow from the origin node i to the destination node j travels only via the path $i \to i \to m \to j$ and $i \to k \to j \to j$, if i and j are hubs, respectively.

In the following subsection, we extend the multiple allocation hub location model of Hamacher et al. (2004) by capturing the finite capacity and the resulting congestion (due to uncertain demand and service times) at the hubs using the service level constraints for the two shipment classes.

2.1. Hub node location model with service level constraints for two shipment classes

As discussed in Section 1, all the shipments arriving at a hub are not homogeneous. Some of them may be regular shipments, while the others may be express shipments with different delivery time requirements. Accordingly, we now extend the model to account for two different shipment classes, indexed by $c \in \{r, e\}$, for the regular (r) and the express (e) delivery options. We redefine the notation accordingly for each customer class $c \in \{e, r\}$. Let λ_{ij}^c be the rate at which requests (number of shipments per unit time) arrive from several shippers for delivery of the shipment class c from the origin node i to the destination node j. Further, let x_{ijkm}^c be the fraction of the shipments for shipment class c from the origin node i to the destination node j that is routed via hubs located at nodes k, m in that order. As discussed in Section 1, we further assume the hubs can be opened only with a finite capacity. For that, let L_k be the set of discrete capacity choices at a candidate hub at node $k \in N$, and let $z_{kl} = 1$ if a hub is opened at node k with the capacity level $l \in L_k$; 0 otherwise. Let μ_{kl} be the capacity (service rate) corresponding to the capacity level l at hub k for both the shipment classes. Since hubs have a finite capacity, the variability in the arrival and service of the shipments at hubs, as discussed in Section 1, creates congestion. To account for this congestion, each hub is modeled as a single server queueing facility, where the mean service rate of hub k is given by $\mu_k = \sum_{l \in L_k} \mu_{kl} z_{kl} : \sum_{l \in L_k} z_{kl} \leq 1$. The capacity at a hub reflects essentially the number of shipments it can serve in a given time period. Shipments within each class are served at a hub on a first-come-first-served (FCFS) basis. However, the express shipments are given preemptive priority in service over the regular shipments.

The congestion at the hubs may cause the shipments to miss their promised delivery times, which may result in penalties, either in the form of a discount, partial refund or an expedited delivery (to avoid any further delay) without additional charge to the customer. Hence, the 3PL service provider sets its own internal target dwell time (also called the sojourn time in the queuing literature) τ^c and a target service level $\beta^c \in (0, 1)$, which is the minimum probability with which a shipment from class c at any hub should be served within τ^c. If we let W_k^c denote the total time spent by any shipment from class c class at hub k, called dwell time, then the service level constraint can be expressed as follows:

$$S_k(\tau^c) = P\{W_k^c \leq \tau^c\} \geq \beta^c \qquad \forall k \in N$$

The objective of the 3PL service provider is to locate hubs with adequate service capacities and select the routes for all the origin-destination pairs via some hubs such that the total network cost is minimized, subject to a separate service level constraint for each shipment class at their consolidation hubs. We refer to this problem as the *Hub Node Location Problem with Two-class Service Level Constraints* (HNLP-TSLC). We define the following

Indices and Sets:		
i, j, k, m	:	Nodes
k, m	:	Hub nodes
l	:	capacity level at hub
c	:	shipment class; $c \in \{e, r\}$.
N	:	Set of all nodes that exchange traffic; $\{i, j, k, m \in N\}$; $N = \{0, 1, 2, \ldots, \lvert N-1 \rvert\}$.
L_k	:	Set of all capacity levels at hub k; $\{l \in L_k\}$; $L_k = \{1, 2, \ldots, \lvert L_k \rvert\}$.
Parameters:		
λ_{ij}^c	:	Mean demand rate (number of shipments per unit time) for the shipment class c from the origin node $i \in N$ to the destination node $j \in N$.
μ_{kl}	:	Capacity (number of shipments per unit time) corresponding to the capacity level l at the hub k.
α	:	Inter-hub shipment discount; $\alpha \in (0, 1)$.
δ	:	Spoke to hub shipment discount; $\delta < \alpha$.
γ	:	Hub to spoke shipment discount; $\gamma < \alpha$.
C_{ij}	:	Transportation cost per unit of direct shipment from the node $i \in N$ to the node $j \in N$.
C_{ijkm}	:	Transportation cost per unit of shipment from the node $i \in N$ to the node $j \in N$ routed via the hubs $k, m \in N$ in that order. $C_{ijkm} = \delta C_{ik} + \alpha C_{km} + \gamma C_{mj}$.
F_{kl}	:	Amortized cost of locating a hub with the capacity level l at hub k.
τ^c	:	Maximum threshold on the dwell time (in queue + in service) for the shipment class c.
β^c	:	Target service level for the shipment class c at a hub.
Variables:		
z_{kl}	:	1 if node k is opened as a hub with capacity level l; 0 otherwise.

(*continued on next page*)

x^c_{ijkm}	:	fraction of the for the shipments from class c from the origin node $i \in N$ to the destination node $j \in N$ that is routed via the hubs located at nodes $k, m \in N$ in that order.
Derived Variables:		
Λ^c_k	:	Rate of arrival (number of shipments per unit time) from class c at hub k.
μ_k	:	Capacity (number of shipments per unit time) installed at the hub k.
W^c_k	:	Dwell time (in queue + in service) for the shipment class c at hub k.
$S^c_k(\tau^c)$	:	Service level achieved for the shipment class c at the hub k, i.e., $P\{W^c_k \le \tau^c\}$.

The resulting mixed integer program (MILP) formulation of HNLP-TSLC is as follows:

[HNLP-TSLC]:

$$\min \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{m\in N}\sum_{c\in\{e,r\}} C_{ijkm}\lambda^c_{ij}x^c_{ijkm} + \sum_{k\in N}\sum_{l\in L_k} F_{kl}z_{kl} \tag{9}$$

$$\text{s.t.} \sum_{k\in N}\sum_{m\in N} x^c_{ijkm} = 1 \qquad \forall i, j \in N, c \in \{e, r\} \tag{10}$$

$$\sum_{m\in N} x^c_{ijkm} + \sum_{m\in N\setminus\{k\}} x^c_{ijmk} \le \sum_{l\in L_k} z_{kl} \qquad \forall i, j, k \in N, c \in \{e, r\} \tag{11}$$

$$\sum_{l\in L_k} z_{kl} \le 1 \qquad \forall k \in N \tag{12}$$

$$\Lambda^c_k \le \sum_{l\in L_k} \mu_{kl} z_{kl} \qquad \forall k \in N, c \in \{e, r\} \tag{13}$$

$$\Lambda^c_k = \sum_i \sum_j \sum_m \lambda^c_{ij} x^c_{ijkm} \qquad \forall k \in N, c \in \{e, r\} \tag{14}$$

$$S^c_k(\tau^c) = P\{W^c_k \le \tau^c\} \ge \beta^c \sum_{l\in L_k} z_{kl} \qquad \forall k \in N, c \in \{e, r\} \tag{15}$$

$$x^c_{ijkm} \ge 0 \qquad \forall i, j, k, m \in N, c \in \{e, r\} \tag{16}$$

$$z_{kl} \in \{0, 1\} \qquad \forall k \in N, l \in L_k \tag{17}$$

Constraints (9)–(11) are counterparts, in a two-class, multi-capacity level setting, of (1)–(3). Constraint set (12) allows a node to be opened as a hub with only one level of capacity. Constraint set (13) ensures the stability of the queueing system at open hubs, where Λ^c_k is the rate of arrival of the (collection) shipments entering hub k directly from the origin nodes, given by (14). Note that Λ_k in (14) captures only the (collection) flows entering hub k directly from the origin node. It does not capture the (transfer) flows entering hub k via another hub. This makes sense in situations where the shipments once processed (e.g., sorted) after collection do not need further processing for distribution (Ebery et al., 2000). However, in situations where the shipments need further processing before distribution, (14) should be modified as Camargo et al. (2009); Marin (2005a):

$$\Lambda^c_k = \sum_i \sum_j \sum_m \lambda^c_{ij} x^c_{ijkm} + \sum_i \sum_j \sum_{m\ne k} \lambda^c_{ij} x^c_{ijmk} \qquad \forall k \in N, c \in \{e, r\} \tag{14-1}$$

Here, the second summation term captures the flows entering hub k only via another hub (transfer shipments). Constraint set (15) are the service level constraints at the hub nodes. The term $\sum_{l\in L_k} z_{kl}$ in the right hand side of (15) ensures that the service level constraints apply only to those nodes that are designated as hubs. The target service level β^c is set based on the importance of the shipment class.

It may be noted HNLP-TSLC does not contain counterparts of the constraint sets (6)–(8), which model the restriction of at most two hubs on any feasible path from an origin node to a destination node in UMAHLP. In the absence of such constraints, a shipment from an origin node i to a destination node j, when both i and j are hubs, may be routed using any of the three different sets of variables: (a) x_{iji} corresponding to the route $i \to i \to i \to j$; (b) x_{ijj} corresponding to the route $i \to j \to j \to j$; and (c) x_{iij} corresponding to the route $i \to i \to j \to j$. All these three sets of variables represent essentially the same route $(i \to j)$. However, (a) and (b) have higher associated costs than (c) since they do not involve any inter-hub discount (α). In this example, the arc from hub i to hub j on the route $i \to i \to j \to j$ with inter-hub discount is called a hub arc, while the same on the routes $i \to i \to i \to j$ and $i \to j \to j \to j$ with no inter-hub discount is referred as "bridge arc". The concept of bridge arcs has been described in detail by Campbell et al. (2005a,b). HNLP-TSLC may prefer (a) if hub i has, while hub j does not have, enough spare capacity to meet the service level constraint. Or, it may prefer (b) if hub j has, while hub i does not have enough capacity to meet the service level constraint. Alternatively, it may route the shipment partly via all the above three routes.

One of the key features of hub networks is the flow consolidation at hub facilities, which leads to a discounted transportation cost on the links. The hub node location models, presented in Section 2.1, have a number of attractive theoretical features. However, the fully interconnected set of hubs (arising as a consequence of the commonly used assumptions of: (a) triangle inequality between the distances, (b) absence of hub arc setup costs, and (c) common inter-hub discount factor (α)) can lead to inconsistencies between the assumed economies of scale and the actual shipments on the different links in the network (Campbell, 2013). In order to overcome this issue, Bryan (1998); O'Kelly & Bryan (1998) and Kimms (2006), among others, incorporated flow dependent transportation cost on the hub arcs, while Campbell et al. (2005a,b) introduced the so-called *hub arc location models*. Note that this may result in a disconnected hub network. In Section 2.2, we present one such hub arc location version of the current problem.

2.2. Hub arc location model with service level constraints for two shipment classes

The basic hub arc location problem seeks to locate discounted hub arcs (whose endpoints are hubs), as opposed to locating hub nodes (which are connected by hub arcs) to allow for a less than fully interconnected set of hubs and prevent the kind of inconsistencies described above. We now present a variant of the hub arc location model that seeks to locate a given number (denoted by q) of hub arcs in the network with service level constraints for two shipment classes. For this, let $y_{km} = 1$ if a hub arc from hub k to m is established, 0 otherwise. The resulting MIP formulation, referred to as *q-Hub Arc Location Problem with Two-class Service Level Constraints* (q-HALP-TSLC), is presented below:

[q-HALP-TSLC]:

$$\min \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{m\in N}\sum_{c\in\{e,r\}} C_{ijkm}\lambda^c_{ij}x^c_{ijkm} + \sum_{k\in N}\sum_{l\in L_k} F_{kl}z_{kl} \tag{18}$$

$$\text{s.t.} \sum_{k\in N}\sum_{m\in N} x^c_{ijkm} = 1 \qquad \forall i, j \in N, c \in \{e, r\} \tag{19}$$

$$\sum_{m\in N} x^c_{ijkm} + \sum_{m\in N\setminus\{k\}} x^c_{ijmk} \le \sum_{l\in L_k} z_{kl} \qquad \forall i, j, k \in N, c \in \{e, r\} \tag{20}$$

$$\sum_{l\in L_k} z_{kl} \le 1 \qquad \forall k \in N \tag{21}$$

$$\Lambda_k^c \le \sum_{l\in L_k} \mu_{kl} z_{kl} \qquad \forall k \in N \tag{22}$$

$$\Lambda_k^c = \sum_i \sum_j \sum_m \lambda_{ij}^c x_{ijkm}^c \qquad \forall k \in N, c \in \{e, r\} \tag{23}$$

$$S_k^c(\tau^c) = P\{W_k^c \le \tau^c\} \ge \beta^c \sum_{l\in L_k} z_{kl} \qquad \forall k \in N, c \in \{e, r\} \tag{24}$$

$$x_{ijkm}^c \le y_{km} \qquad \forall i, j, k, m, \in N, c \in \{e, r\} \tag{25}$$

$$y_{km} \le \sum_{l\in L_k} z_{kl} \qquad \forall k \in N, m \in N | m \ne k \tag{26}$$

$$y_{km} \le \sum_{l\in L_k} z_{ml} \qquad \forall k \in N, m \in N | m \ne k \tag{27}$$

$$\sum_{k\in N} \sum_{m\in N | m\ne k} y_{km} = q \tag{28}$$

$$y_{km} \in \{0, 1\} \qquad \forall k, m \in N \tag{29}$$

$$x_{ijkm}^c \ge 0 \qquad \forall i, j, k, m \in N, c \in \{e, r\} \tag{30}$$

$$z_{kl} \in \{0, 1\} \qquad \forall k \in N, l \in L_k \tag{31}$$

We define the additional constraints introduced in this subsection. Constraint set (25) ensures that any shipment via hubs k and m is not routed unless a hub arc is established from node k to node m. Constraint sets (26) and (27) together ensure that a hub arc from node k to node m can only be established if both k and m are open as hubs. Constraint set (28) restricts the number of hub arcs to be established to q. Constraint (29) is the set of binary constraints related to hub arc location variables.

The formulations HNLP-TSLC or q-HALP-TSLC are generic in that they provide the following special cases with the below-suggested modifications to the model:

$\lambda_{ij}^r =$ $0, \beta^r =$ 0	:	Capacitated Model for Single Class with Service Level Constraint ($P\{W_k^e \le \beta^e\}$).
$\lambda_{ij}^e =$ $0, \beta^e =$ 0	:	Capacitated Model for Single Class with Service Level Constraint ($P\{W_k^r \le \beta^r\}$).
$\beta^e =$ $0, \beta^r =$ 0	:	Capacitated Model for Single Class without Service Level Constraint (since there is no service level requirement for either class, the two classes are identical).

If we assume that the rate of flows for the shipment class c between different origin node-destination node pairs (i, j) are independent random variables that follow a Poisson process with mean λ_{ij}^c, then the aggregate flow rate through hub k also follows a Poisson process with a mean given by (14).

Further, if the service times at the hub follow an exponential distribution, then there are two approaches to modeling the queuing system at a hub. One approach is to model the hub as a single-server queuing system with flexible service capacity μ, which can be adjusted either continuously or in discrete steps. The second approach is to model the hub as a queuing system with multiple parallel servers, each with a given single capacity level μ. In this case, the decision variable is the appropriate number of servers to be installed at the hub. In this paper, we adopt the former approach and model each hub facility as a preemptive priority (to model the preferential treatment for express shipments) single server with multiple capacity levels, from which one capacity level is to be selected if the hub is opened. We take this approach primarily for tractability of the resulting model, given that we have priority queues as a result of two shipment classes. However, a single server model may still be a good approximation of a multi-sever facility if the utilization of the service facility is reasonably high. This is because under reasonably high system utilization, a system with s parallel servers, each with capacity μ, is known to have a performance similar to a single server with capacity $s\mu$. The use of an M/M/1 queue to model a hub facility in hub networks is also supported by the literature summarized in Table 1.

Since the sojourn time distribution $S_k^e(\tau^e) = P\{W_k^e \le \tau^e\}$ for high priority (express) customers in a preemptive priority M/M/1 queue is known to be exponential, its service level constraint ((15) in HNLP-TSLC or (24) in q-HALP-TSLC) can be specified as (Gross & Harris, 1998):

$$\sum_{l\in L_k} \mu_{kl} z_{kl} - \Lambda_k^e \ge \frac{-\ln(1-\beta^e)}{\tau^e} \sum_{l\in L_k} z_{kl} \tag{32}$$

However, such an analytical characterization of the sojourn time distribution $S_k^r(\tau^r) = P\{W_k^r \le \tau^r\}$ for low priority (regular) customers is not known (Abate & Whitt, 1997). This makes HNLP-TSLC or q-HALP-TSLC challenging to solve. In the following section, we discuss how we address service level constraints for regular customers (corresponding to (15) in HNLP-TSLC or (24) in q-HALP-TSLC).

3. Solution methodology

The absence of an analytical characterization of the service level constraint (15) in HNLP-TSLC (or (24) in q-HALP-TSLC) for the regular customers (shipments) makes the two models challenging to solve. While the Laplace transform of the sojourn time distribution $S_k^r(\tau^r)$, appearing in (15), and its first few moments are well known (Stephan, 1958), the distribution itself is somewhat complicated and requires numerical computation of the inverse Laplace transform, thereby preventing its analytical characterization (Jayaswal et al., 2011; Jayaswal & Vidyarthi, 2017). There are approximations proposed in the literature for the sojourn time distribution. However, they are very complex and often not sufficiently accurate (Abate & Whitt, 1997). Moreover, the choice of appropriate approximation to be used depends on Λ_k^e and Λ_k^r, which can only be determined endogenously, and are not known in advance in our model.

Although the exact form of $S_k^r(\tau^r)$ in constraint (15) in HNLP-TSLC (or (24) in q-HALP-TSLC) is unknown, we exploit its special structure, determined numerically using the matrix geometric method. Plots of $S_k^r(\tau^r)$ vs. $(\Lambda_k^e, \Lambda_k^r)$, $S_k^r(\tau^r)$ vs. (Λ_k^e, μ_k) and $S_k^r(\tau^r)$ vs. (Λ_k^r, μ_k) are shown in Fig. 1. These plots suggest that $S_k^r(\tau^r)$ is jointly concave in $(\Lambda_k^e, \Lambda_k^r)$, in (Λ_k^e, μ_k), and also in (Λ_k^r, μ_k). However, this does not necessarily show the joint concavity of $S_k^r(\tau^r)$ in $(\Lambda_k^e, \Lambda_k^r, \mu_k)$. We, therefore, integrate into our solution method a mechanism (see Section 3.3) to ensure that the concavity assumption is not violated. Assuming $S_k^r(\tau^r)$ is concave, it can be approximated by a set of tangent hyperplanes at various points $((\Lambda_k^e)^p, (\Lambda_k^r)^p, (\mu_k)^p)$, $\forall\ p \in P$:

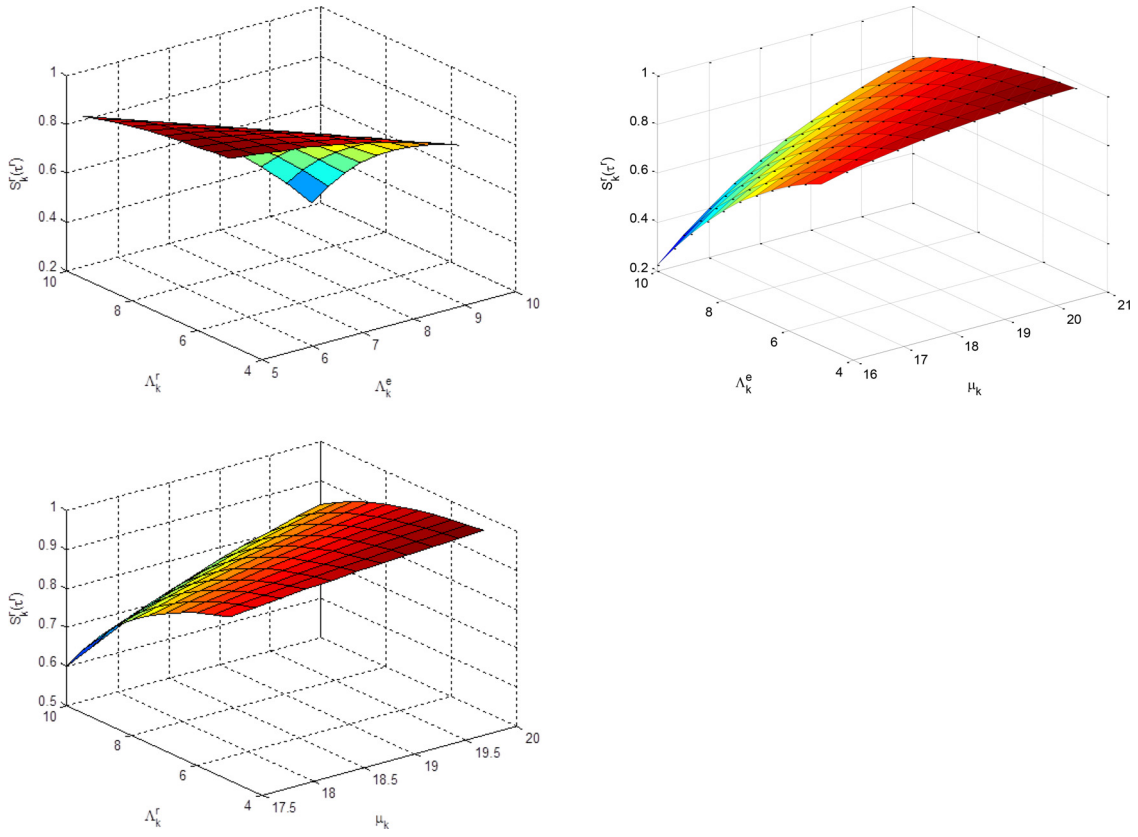

Fig. 1. Service level for the regular shipments at hub k vs. demands for the regular and express shipments and hub capacity.

$$S_k^r(\tau^r) = \min_{p\in P}\left\{(S_k^r(\tau^r))^p + \left(\Lambda_k^e - (\Lambda_k^e)^p\right)\left(\frac{\partial(S_k^r(\tau^r))}{\partial\Lambda_k^e}\right)^p + \left(\Lambda_k^r - (\Lambda_k^r)\right)^p\left(\frac{\partial(S_k^r(\tau^r)}{\partial\Lambda_k^r}\right) + (\mu_k - (\mu_k)^p)\left(\frac{\partial(S_k^r(\tau^r))}{\partial\mu_k}\right)^p\right\},$$

where $(S_k^r(\tau^r))^p$ denotes the value of $S_k^r(\tau^r)$ at a fixed point $((\Lambda_k^e)^p,\ (\Lambda_k^r)^p,\ (\mu_k)^p)$, and $\left(\frac{\partial(S_k^r(\tau^r))}{\partial\Lambda_k^e}\right)^p$, $\left(\frac{\partial(S_k^r(\tau^r))}{\partial\Lambda_k^r}\right)^p$, and $\left(\frac{\partial(S_k^r(\tau^r))}{\partial\mu_k}\right)^p$ are the gradients of $S_k^r(\tau^r)$ at $((\Lambda_k^e)^p,\ (\Lambda_k^r)^p,\ (\mu_k)^p)$. Constraint (15) (or (24)) for $c = r$ can thus be replaced by the following set of linear constraints:

$$(S_k^r(\tau^r))^p + \left(\Lambda_k^e - (\Lambda_k^e)^p\right)\left(\frac{\partial(S_k^r(\tau^r))}{\partial\Lambda_k^e}\right)^p + \left(\Lambda_k^r - (\Lambda_k^r)\right)^p\left(\frac{\partial(S_k^r(\tau^r)}{\partial\Lambda_k^r}\right) + (\mu_k - (\mu_k)^p)\left(\frac{\partial(S_k^r(\tau^r))}{\partial\mu_k}\right)^p \geq \beta^r \quad \forall p \in P \tag{33}$$

Replacing (15) (or (24)) for $c = r$ by (33) results in a finite but a large number of constraints, which is amenable to the cutting plane method.

We use the matrix geometric method to numerically evaluate $(S_k^r(\tau^r))^p$ at a given point $((\Lambda_k^e)^p,\ (\Lambda_k^r)^p,\ (\mu_k)^p)$. We refer the readers to Neuts (1981) for details of the matrix geometric method. The use of the matrix geometric method yields explicit recursive formulas for the joint stationary probabilities, which can provide significant computational improvements over the transform techniques (Miller, 1981). Moreover, it gives exact solutions, in contrast to discrete event simulation, which is another alternative method to evaluate $S_k^r(\tau^r)$ that at best gives point estimates. The matrix geometric method is also computationally efficient compared to simulation. This is important in solving [*CMAHLP* – *TSLC*], which requires repeated evaluation of $(S_k^r(\tau^r))^p$ for various open hubs k at various solutions points p. Once $S_k^r(\tau^r)$ is evaluated at a point $((\Lambda_k^e)^p,\ (\Lambda_k^r)^p,\ (\mu_k)^p)$, its gradients are obtained using the *finite difference method* (described in Section 3.2). The gradients are used to generate cuts of the form (33), 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 3.3.

3.1. The matrix geometric method

3.1.1. The joint stationary queue length distribution at the hub

If we define $N_k^e(t)$ and $N_k^r(t)$ as state variables representing the number of express (high priority) and regular (low priority) shipments at hub k at time t, then $\{\mathbf{N_k}(t)\} := \{N_k^r(t), N_k^e(t), t \geq 0\}$ is a continuous-time two-dimensional Markov chain with state space $\{\mathbf{n_k} = (n_k^r, n_k^e)\}$. The key idea we employ here is that $\{\mathbf{N_k}(t)\}$ is a *quasi-birth-and-death* (QBD) process, which allows us to develop a matrix geometric solution for the joint distribution of the number of shipments of each class at hub k. A simple implementation of the matrix geometric method, however, requires the number of states in the QBD process to be finite. For this, we treat the queue length of express shipments (including the one in service) to be of finite size M, but of size large enough for the desired accuracy of our results. Since express shipments are always served in priority over regular shipments, it is reasonable to assume that its queue size will always be bounded by some large number.

In the Markov process $\{\mathbf{N_k}(t)\}$, a transition can occur only if a shipment of either class arrives or is served at hub k. The possible transitions are:

From	To	Rate	Condition
(n_k^r, n_k^e)	(n_k^r, n_k^e+1)	Λ_k^e	for $n_k^r \geq 0,\ 0 \leq n_k^e < M$
(n_k^r, n_k^e)	(n_k^r+1, n_k^e)	Λ_k^r	for $n_k^r \geq 0,\ 0 \leq n_k^e \leq M$
(n_k^r, n_k^e)	(n_k^r, n_k^e-1)	μ_k	for $n_k^r \geq 0,\ 0 \leq n_k^e \leq M$
(n_k^r, n_k^e)	(n_k^r-1, n_k^e)	μ_k	for $n_k^r > 0,\ n_k^e = 0$

The infinitesimal generator Q associated with our system description is thus block-tridiagonal:

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where B_0, A_0, A_1, A_2 are square matrices of order $M+1$, as given in Appendix A. We denote $\mathbf{x}$ as the stationary probability vector of $\{\mathbf{N_k}(t)\}$:

$$\mathbf{x} = [x_{00}, x_{01}, \ldots, x_{0M}, x_{10}, x_{11}, \ldots, x_{1M}, \ldots, \ldots, x_{i0}, x_{i1}, \ldots, x_{iM}, \ldots, \ldots]$$

The vector $\mathbf{x}$ can be partitioned by levels into sub vectors $\mathbf{x}_i$, $i \geq 0$, where $\mathbf{x}_i = [x_{i0}, x_{i1}, \ldots, x_{iM}]$ is the stationary probability of states in level i ($n_k^r = i$). Thus, $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \ldots]$. $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form by the following standard relations (Latouche & Ramaswami, 1999; Neuts, 1981):

$$\mathbf{x}Q = \mathbf{0}; \quad \mathbf{x}_{i+1} = \mathbf{x}_i R$$

where, $\mathbf{0}$ is a row vector of zeros of appropriate size, and R is the minimal non-negative solution to the matrix quadratic equation:

$$A_0 + RA_1 + R^2A_2 = \mathbf{0}$$

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

$$R(0) = 0\ ; \quad R(n+1) = -\left[A_0 + R^2(n)A_2\right]A_1^{-1}$$

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

$$\mathbf{x}_0(B_0 + RA_2) = \mathbf{0}$$

subject to the normalization equation:

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

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

3.1.2. Estimation of $S_k^r(\tau^r)$

The dwell (sojourn) time W_k^r of a regular shipment at hub k is the time between its arrival to hub k till it completes service at that hub. It may be preempted by one or more express shipments for service. So it is difficult to characterize the distribution $S_k^r(\cdot)$. Ramaswami & Lucantoni (1985) present an efficient algorithm based on *uniformization* to derive the complementary distribution of waiting times in phase-type and QBD processes (see Gross & Harris, 1998, for details on uniformization (also referred to as randomization)). Jayaswal et al. (2011); Jayaswal & Vidyarthi (2017) adapt their algorithm to derive $S_k^r(\cdot)$, the distribution of the sojourn time (waiting time plus the time in service) of low priority (regular) customers, which we adopt in this paper.

Consider a tagged regular shipment entering the system. The time spent by the tagged shipment depends on the number of shipments of either class already present in the system ahead of it, and also on the number of subsequent express arrivals before it completes its service. All subsequent regular arrivals, however, have no influence on its time spent in the system. The tagged shipment's time in the system is, therefore, simply the time until absorption in a modified Markov process $\{\tilde{\mathbf{N}}_\mathbf{k}(t)\}$, obtained by setting $\Lambda_k^r = 0$. Consequently, matrix $\tilde{A_0}$, representing transitions to a higher level, becomes a zero matrix. We define an *absorbing* state, call it state $0'$, as the state in which the tagged shipment has finished its service. The infinitesimal generator for this process can be represented as:

$$\tilde{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \ldots \\ b_0 & \tilde{B_0} & 0 & & & \\ 0 & A_2 & \tilde{A_1} & 0 & & \\ 0 & & A_2 & \tilde{A_1} & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{pmatrix}$$

where, $\tilde{B_0} = B_0 + A_0$; $\tilde{A_1} = A_1 + A_0$; and $b_0 = [\mu_k \quad 0 \quad \ldots \quad 0]^T_{M+1}$. The first row and column in $\tilde{Q}$ corresponds to the absorbing state $\acute{0}$. The time spent in system by the tagged shipment, which is the time until absorption in the modified Markov process with rate matrix $\tilde{Q}$, depends on the arrival rates Λ_k^e and Λ_k^r and the capacity μ_k at hub k. For a given point p (corresponding to arrival rates $(\Lambda_k^e)^p$, $(\Lambda_k^r)^p$ and capacity $(\mu_k)^p$ at hub k) in the solution space, the distribution of the time spent by a regular shipment at hub k is $(S_k^r(\tau^r))^p = 1 - \overline{(S_k^r(\tau^r))^p}$, where $\overline{(S_k^r(\tau^r))^p}$ is the stationary probability that a regular shipment spends more than τ^r units of time at hub k. Further, let $\overline{(S_{kij}^r(\tau^r))^p}$ denote the conditional probability that a tagged shipment, which finds i regular and j express shipments ahead of it, spends a time exceeding τ^r at hub k. The probability that a tagged shipment finds i regular and j express shipments ahead of it is given, using the PASTA property, by $\mathbf{x}_{ij}$. $\overline{(S_k^r(\tau^r))^p}$ can thus be expressed as:

$$\overline{(S_k^r(\tau^r))^p} = \sum_{i=0}^{\infty} \sum_{j=0}^{M} \mathbf{x}_{ij} \overline{(S_{kij}^r(\tau^r))^p}$$

We know that transitions out of state ij ($n_k^r = i, n_k^e = j$) in the Markov Process $\{\tilde{\mathbf{N}}_\mathbf{k}(t)\}$ occur according to a Poisson process with rate γ_{ij}, where $\gamma_{0j} = (-\tilde{B_0})_{jj}$ and $\gamma_{ij} = (-\tilde{A_1})_{jj}$ for $i \geq 1$. Thus, the probability that n transitions are generated out of state ij in time τ^r is $e^{-\gamma_{ij}\tau^r} \frac{(\gamma_{ij}\tau^r)^n}{n!}$. Suppose the tagged shipment finds i regular and j express shipments ahead of it. Then, for its dwell time at hub k to exceed τ^r, at most i of the n transitions out of state ij may correspond to lower levels (i.e., service completions of regular shipments). Therefore,

$$\overline{(S_{kij}^r(\tau^r))^p} = \sum_{n=0}^{\infty} e^{-\gamma_{ij}\tau^r} \frac{(\gamma_{ij}\tau^r)^n}{n!} \sum_{v=0}^{i} P_{ij,v}^{(n)}, \qquad i \geq 0$$

where, $P_{ij,v}^{(n)}$ is the conditional probability, given that the Markov Process $\{\tilde{\mathbf{N}}_\mathbf{k}(t)\}$ has made n transitions out of state ij, that v of those transitions correspond to lower levels (i.e., service completions of regular shipments). Sojourn time distribution for regular shipments can, therefore, be expressed as:

$$(S_k^r(\tau^r))^p = 1 - \overline{(S_k^r(\tau^r))^p} = 1 - \sum_{i=0}^{\infty} \sum_{j=0}^{M} \mathbf{x}_{ij} \sum_{n=0}^{\infty} e^{-\gamma_{ij}\tau^r} \frac{(\gamma_{ij}\tau^r)^n}{n!} \sum_{v=0}^{i} P_{ij,v}^{(n)}$$

However, $(S_k^r(\tau^r))^p$ can be computed more conveniently via uniformization (Gross & Harris, 1998). We uniformize the Markov process $\{\tilde{\mathbf{N}}_\mathbf{k}(t)\}$ with a Poisson process with rate γ, where

$$\gamma = \max_{0 \leq j \leq M} (-\tilde{A_1})_{jj} = \max_{0 \leq j \leq M} -(A_0 + A_1)_{jj}$$

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

$$\hat{Q} = \frac{1}{\gamma}\tilde{Q} + I = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots \\ \hat{b_0} & \hat{B_0} & 0 & & & \\ 0 & \hat{A_2} & \hat{A_1} & 0 & & \\ 0 & & \hat{A_2} & \hat{A_1} & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{pmatrix}$$

where $\hat{A_2} = \frac{A_2}{\gamma}$, $\hat{A_1} = \frac{\tilde{A_1}}{\gamma} + I$, $\hat{b_0} = \frac{b_0}{\gamma}$. In this uniformized process, transitions from any state are generated at a rate γ. Further, let $\overline{(S^r_{ki}(\tau^r))^p}$ denote the conditional probability that a tagged shipment, which finds i regular shipments ahead of it, spends a time exceeding τ^r at hub k. The probability that a tagged shipment finds i regular shipments is given, using the PASTA property, by $\mathbf{x}_i = \mathbf{x}_0 R^i$. $\overline{(S^r_k(\tau^r))^p}$ can be expressed as:

$$\overline{(S^r_k(\tau^r))^p} = \sum_{i=0}^{\infty} \mathbf{x}_i \overline{(S^r_{ki}(\tau^r))^p} \mathbf{1} \tag{34}$$

The probability that n Poisson events are generated in time τ^r is $e^{-\gamma\tau^r}\frac{(\gamma\tau^r)^n}{n!}$. Suppose the tagged shipment finds i regular shipments ahead of it. Then, for its dwell time at hub k to exceed τ^r, at most i of the n Poisson points may correspond to transitions to lower levels (i.e., service completions of regular shipments). Therefore,

$$\overline{(S^r_{ki}(\tau^r))^p} = \sum_{n=0}^{\infty} e^{-\gamma\tau^r}\frac{(\gamma\tau^r)^n}{n!}\sum_{v=0}^{i} G_v^{(n)}\mathbf{1}, \qquad i \geq 0 \tag{35}$$

where, $G_v^{(n)}$ 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 (i.e., service completions of regular shipments). Substituting the expression for $\overline{(S^r_{ki}(\tau^r))^p}$ from (35) into (34), we obtain:

$$\overline{(S^r_k(\tau^r))^p} = \sum_{n=0}^{\infty} d_n e^{-\gamma\tau^r}\frac{(\gamma\tau^r)^n}{n!} \tag{36}$$

where, d_n is given by:

$$d_n = \sum_{i=0}^{\infty} \mathbf{x}_0 R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{1}, \qquad n \geq 0 \tag{37}$$

Further algebraic manipulations of (37), as detailed in Appendix B, gives:

$$(S^r_k(\tau^r))^p = 1 - \overline{(S^r_k(\tau^r))^p} = 1 - \sum_{n=0}^{\infty} e^{-\gamma\tau^r}\frac{(\gamma\tau^r)^n}{n!}\mathbf{x}_0(I-R)^{-1}H_n\mathbf{1} \tag{38}$$

where, $H_n = \sum_{v=0}^{n} R^v G_v^{(n)}$, which can be computed recursively as:

$$H_{n+1} = H_n\hat{A}_1 + RH_n\hat{A}_2; \quad H_0 = I$$

where, I is an identity matrix of size $M+1$. Therefore, for given arrival rates $((\Lambda^e_k)^p, (\Lambda^r_k)^p)$ and capacity $((\mu_k)^p)$ at hub k, $S^r_k(\tau^r)$ in (16) can be computed using (38).

3.2. Estimation of the gradient of $S^r_k(\tau^r)$

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

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\Lambda^e_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p + d\Lambda^e_k, (\Lambda^r_k)^p, (\mu_k)^p)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p - d\Lambda^e_k, (\Lambda^r_k)^p, (\mu_k)^p)}}{2d\Lambda^e_k}$$

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\Lambda^r_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p + d\Lambda^r_k, (\mu_k)^p)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p - d\Lambda^r_k, (\mu_k)^p)}}{2d\Lambda^r_k}$$

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\mu_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p + d\mu_k)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p - d\mu_k)}}{2d\mu_k}$$

where $d\Lambda^e_k$, $d\Lambda^r_k$ and $d\mu_k$ (referred to as step sizes) are infinitesimal changes in the respective variables. However, when $(\Lambda^e_k)^p < d\Lambda^e_k$, $(\Lambda^r_k)^p < d\Lambda^r_k$ or $(\mu_k)^p < d\mu_k$, then $(\Lambda^e_k)^p - d\Lambda^e_k < 0$, $(\Lambda^r_k)^p - d\Lambda^r_k < 0$ or $(\mu_k)^p - d\mu_k < 0$, in which case the computed service level function (appearing in the numerator of the gradient equation) does not have any physical meaning. To avoid such odd situations, the corresponding gradient in such a case is estimated using the forward difference method as:

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\Lambda^e_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p + d\Lambda^e_k, (\Lambda^r_k)^p, (\mu_k)^p)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p)}}{d\Lambda^e_k}$$

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\Lambda^r_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p + d\Lambda^r_k, (\mu_k)^p)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p)}}{d\Lambda^r_k}$$

$$\left(\frac{\partial(S^r_k(\tau^r))}{\partial\mu_k}\right)^p = \frac{(S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p + d\mu_k)} - (S^r_k(\tau^r))^{((\Lambda^e_k)^p, (\Lambda^r_k)^p, (\mu_k)^p)}}{d\mu_k}$$

3.3. The cutting plane algorithm

The cutting plane algorithm to solve HNLP-TSLC or q-HALP-TSLC is given below. The algorithm differs from the traditional description 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 success of the cutting plane algorithm relies on the concavity of $S^r_k(\tau^r)$. We have demonstrated, using computational results obtained by the matrix geometric method, that $S^r_k(\tau^r)$ is concave in $(\Lambda^e_k, \Lambda^r_k)$ and separately concave in μ_k. However, it is difficult to establish the joint concavity of $S^r_k(\tau^r)$ in $(\Lambda^e_k, \Lambda^r_k, \mu_k)$. If the concavity assumption is violated, then the algorithm may cut off parts of the feasible region and terminate with a solution that is suboptimal. We conduct a test to ensure the concavity assumption is not violated. This is done by ensuring that a new point, visited by the cutting plane algorithm after each iteration, lies below all the previously defined cuts, and that all previous points lie below the newly added cut. The test, however, cannot ensure that $S^r_k(\tau^r)$ is concave unless it examines all the points in the feasible region. Still, it does help ensure that the concavity assumption is not violated at least in the region visited by the algorithm. We used this test in our numerical experiments, which did ensure that the concavity assumption was not violated for any of the instances studied, at least in the region visited by the algorithm. Details of the test can be found in Atlason et al. (2004).

4. Computational study

In this section, we present our computational experiments on test instances described in Section 4.1, followed by the results and

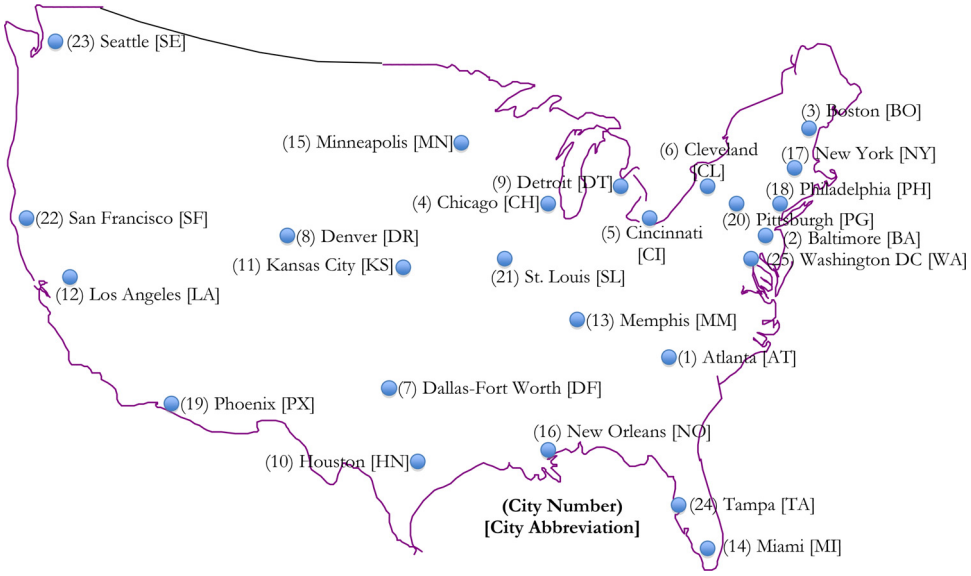

Fig. 2. Cities in CAB data.

their analysis in Section 4.2. All the computational experiments are performed on a personal computer with Intel(R) Core(TM) i7-7000 CPU @ 3.60 gigahertz processor with 16 gigabyte RAM and 64-bit Windows operating system. The cutting-plane algorithm, described in Algorithm 1, is coded in C++. The default MILP solver of Cplex 22.1.0.0 is used to solve the MILPs arising in step 3 of the algorithm.

Algorithm 1 Cutting plane algorithm.

1: $P \leftarrow \Phi$.
2: **repeat**
3: Solve HNLP-TSLC(P) or q-HALP-TSLC(P) to obtain x^c_{ijkm} $\forall c \in \{e, r\}$ and z_{kl} $\forall k \in N, l \in L_k$.
4: Obtain Λ^e_k and Λ^r_k using (14) and $\mu_k = \sum_{l \in L_k} \mu_{kl} z_{kl}$ $\forall k \in \{N : \sum_{l \in L_k} z_{kl} = 1\}$. $p \leftarrow \left\{\left(\Lambda^e_k, \Lambda^r_k, \mu_k\right)\right\}_{k \in N: \sum_{l \in L_k} z_{kl} = 1}$
5: Obtain $S^r_k(\tau^r)$ using (38) $\forall k \in \{N : \sum_{l \in L_k} z_{kl} = 1\}$.
6: **if** $S^r_k(\tau^r) \geq \beta^r$ $\forall k \in \{N : \sum_{l \in L_k} z_{kl} = 1\}$ **then**
7: Stop.
8: **else**
9: Obtain cuts of the form (33) $\forall k \in \{N : \sum_{l \in L_k} z_{kl} = 1\}$.
10: $P \leftarrow P \cup \{p\}$.
11: **end if**
12: **until** $S^r_k(\tau^r) < \beta^r$ for any $k \in \{N : \sum_{l \in L_k} z_{kl} = 1\}$.

4.1. Test instances

We report our computational results for problem instances based on the US Civil Aeronautics Board (CAB) data for $|N| = 25$ cities (cities in CAB data are indicated in Fig. 2). However, the data set does not contain hub capacities (μ_{kl}) and the associated fixed costs (F_{kl}), required for our problem. So, we generate these additional data using the data generation scheme described below.

Flow (λ_{ij}) between each node pair (i, j) provided in the CAB data set are scaled such that $TF = \sum_{i \in N} \sum_{j \in N} \lambda_{ij} = 2.0$, where TF is the total flow in the network. Further, to model randomness in the flows, we treat the scaled flow between a pair of nodes obtained from the CAB data as the mean (per hour) of the Poisson flows. We set 3 potential capacity levels for any hub $k \in N$ as $\mu_{kl} = \{1.0, 2.0, 3.0\}$ (per hour) for $l \in L_k = \{1, 2, 3\}$, corresponding to low, medium, and high capacity levels. Capacity levels are so selected to represent the case where the medium capacity level at a hub is sufficient to handle 100% of the total flow rate originating in the network, while low and high capacity levels at a hub are sufficient to handle 50% and 150%, respectively. The fixed cost of opening a hub with a capacity μ_{kl} is generated using the function: $F_{kl} = 200(\mu_{kl})^a$, where $a \in \{0.50, 0.75\}$ represents the economy of scale in installing a capacity at a hub. Inter-hub flow discount factor α is selected from the set $\{0.50, 0.75\}$, while the spoke to hub discount factor (δ) and the hub-to-spoke discount factor (γ) are set to 1.0. The composition of the express (e) and regular (r) shipments is represented as: (f^e, f^r), where $f^e \in \{0, 0.25, 0.50, 0.75, 1.0\}$ and $f^r = 1 - f^e$ are the fractions of the total shipments between any pair of nodes that are express and regular, respectively. Thus, $\lambda^e_{ij} = f^e \times \lambda_{ij}$, $\lambda^r_{ij} = f^r \times \lambda_{ij}$ $\forall i, j \in N$. The maximum thresholds on the dwell times at any consolidation hub k for the express and the regular shipments are set (in hours) as $(\tau^e = 6.0,\ \tau^r = 10.0)$. The target service levels (β^e, β^r) at any consolidation hub k for the express and the regular shipments are selected from the set {(90.0%, 90.0%), (95.0%, 95.0%), (98.0%, 98.0%)}.

4.2. Results and discussion

The results of our computational experiments are summarized inTables 2–5 for the Hub Node Location Model. For the Hub Arc Location Model, the results are summarized in Tables 6–9 in Appendix C. In the following, we seek answers to the following important questions:

- How do the service level requirements at the hubs affect the hub network?
- How costly is it to provide a given level of service (β^e, β^r) at hubs? How does this cost vary with different parameters like (β^e, β^r), (f^e, f^r), a, α, and τ^c?

We use the results from our numerical experiments to answer these questions. Important patterns emerging from our results are reported as observations.

4.2.1. Effect of service level requirements on hub network

Tables 2 –5 present comparisons of the results for HNLP-TSLC for different combinations of a, α and (f^e, f^r), corresponding to

Table 2
Effect of service level constraints: hub node location model for $|N| = 25$, $a = 0.50$.

Parameters			Results without service level constraints						Results with service level constraints									
α	f^e	f^r	Hub (capacity level)			Cost	Iter.	CPU	β^e	β^r	Hub (capacity level)				Cost	Iter.	CPU	CoSQ%
0.5	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	6.82	–	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2428.57	1	26.65	0.64
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	17.63	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2447.7	5	162.51	1.43
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.9	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2460.91	6	177.82	1.98
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	16.97	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2474.44	5	141.06	2.54
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	7.17	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2448	1	37.92	1.44
	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	6.88	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2430.77	1	19.09	0.73
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	16.86	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2463.79	6	158.1	2.10
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.68	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2499.7	7	181.11	3.59
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	17.01	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2517.12	4	116.21	4.31
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	6.49	0.95	–	AT(1)	CH(1)	LA(1)	NY(2)	2513.27	1	66.83	4.15
	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	6.56	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2451.15	1	30.95	1.57
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	16.78	0.98	0.98	AT(1)	CH(1)	LA(1)	PH(2)	2523.96	5	124.08	4.59
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.73	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2523.01	4	162.21	4.55
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	16.77	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2554.72	4	404.48	5.87
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	6.41	0.98	–	CH(1)	LA(1)	MM(1)	PH(2)	2553.64	1	99.09	5.82
0.75	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.75	–	0.9	CH(1)	LA(1)	NY(1)		2658.86	1	74.02	1.98
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.83	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2666.97	5	134.01	2.29
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.41	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2673.77	5	108.18	2.55
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	16.38	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2709.23	5	123.19	3.91
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.89	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2679.54	1	110.08	2.77
	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.43	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2667.85	1	147.65	2.33
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.63	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2675.91	5	113.71	2.63
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.27	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2724.11	6	142.96	4.48
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	15.25	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2756.4	4	307.3	5.72
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.44	0.95	–	BA(2)	LA(1)	SL(1)		2717.2	1	85.55	4.22
	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.61	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2682.67	1	98.01	2.89
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.71	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2749.8	6	168.5	5.47
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.39	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2761.03	4	482.41	5.90
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	15.44	0.98	0.98	BA(2)	LA(1)	SL(2)		2787.9	3	374.44	6.93
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.51	0.98	–	BA(2)	LA(1)	SL(1)		2771.57	1	132.01	6.30

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6; CPU = Computation Time (in seconds); CoSQ = Cost of Service Quality.

Table 3
Effect of service level constraints: hub node location model for $|N| = 25$, $a = 0.75$.

Parameters			Results without service level constraints						Results with service level constraints									
α	f^e	f^r	Hub (capacity level)			Cost	Iter.	CPU	β^e	β^r	Hub (capacity level)				Cost	Iter.	CPU	CoSQ%
0.5	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	7.17	–	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2428.57	1	24.08	0.64
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	17.34	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2447.7	5	176.17	1.43
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.95	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2460.91	6	164.82	1.98
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	16.66	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2474.44	5	141.44	2.54
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	6.54	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2448	1	28.1	1.44
	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	6.64	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2430.77	1	15.53	0.73
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	16.79	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2463.79	6	171.56	2.10
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.6	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2499.7	7	182.93	3.59
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	16.95	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2564.81	5	219.19	6.28
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	6.49	0.95	–	CH(1)	DF(1)	LA(1)	NY(1)	2549.51	1	190.66	5.65
	0	1	CH(1)	LA(1)	PH(1)	2413.15	1	6.58	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2451.15	1	29.55	1.57
	0.25	0.75	CH(1)	LA(1)	PH(1)	2413.15	1	16.91	0.98	0.98	AT(1)	CH(1)	LA(1)	PH(1)	2543.75	7	191.79	5.41
	0.5	0.5	CH(1)	LA(1)	PH(1)	2413.15	1	12.9	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2573.52	6	189.25	6.65
	0.75	0.25	CH(1)	LA(1)	PH(1)	2413.15	1	17.13	0.98	0.98	AT(1)	CH(1)	DF(1)	LA(1)	2706.56	5	736.86	12.16
	1	0	CH(1)	LA(1)	PH(1)	2413.15	1	6.49	0.98	–	CH(1)	LA(1)	MM(1)	PH(2)	2607.15	1	126.49	8.04
0.75	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.65	–	0.9	CH(1)	LA(1)	NY(1)		2658.86	1	94.35	1.98
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.72	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2666.97	5	157.46	2.29
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.3	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2673.77	5	149.07	2.55
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	17.55	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2709.23	5	209.02	3.91
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.79	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2679.54	1	70.35	2.77
	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.42	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2667.85	1	87.11	2.33
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.45	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2675.91	5	148.54	2.63
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.38	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2724.11	6	150.78	4.48
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	15.45	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2809.92	4	370.29	7.77
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.41	0.95	–	BA(2)	LA(1)	SL(1)		2770.72	1	183.08	6.27
	0	1	CH(1)	LA(1)	NY(1)	2607.22	1	6.42	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2682.67	1	67.99	2.89
	0.25	0.75	CH(1)	LA(1)	NY(1)	2607.22	1	15.72	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2762.28	6	154.02	5.95
	0.5	0.5	CH(1)	LA(1)	NY(1)	2607.22	1	12.32	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2814.55	4	654.88	7.95
	0.75	0.25	CH(1)	LA(1)	NY(1)	2607.22	1	15.43	0.98	0.98	BA(2)	CH(1)	LA(1)	SL(1)	2868.98	5	848.27	10.04
	1	0	CH(1)	LA(1)	NY(1)	2607.22	1	6.52	0.98	–	BA(2)	LA(1)	SL(1)		2825.09	1	141.02	8.36

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6; CPU = Computation Time (in seconds); CoSQ = Cost of Service Quality.

Table 4
Service levels at the hubs: hub node location model for $|N| = 25$, $a = 0.50$.

Parameters			Without service level constraints					With service level constraints			
α	f^e	f^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			β^e	β^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			
0.5	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.9	AT(–, 99.82)	CH(–, 99.04)	LA(–, 99.87)	NY(–, 90.55)
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.9	0.9	AT(99.57,99.49)	CH(99.38,96.41)	LA(99.59,99.64)	PH(99.31,90.00)
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.9	0.9	AT(99.19,98.59)	CH(98.21,90.00)	LA(99.33,99.17)	PH(98.44,89.99)
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.9	0.9	AT(97.27,92.57)	CH(96.89,89.99)	LA(98.76,98.02)	NY(97.08,90.00)
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.9	–	AT(96.65, –)	CH(90.00, –)	LA(98.16, –)	NY(90.00, -)
	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.95	AT(–, 99.81)	CH(–, 98.30)	LA(–, 99.87)	NY(–, 95.00)
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.95	0.95	AT(99.49,99.17)	CH(99.34,95.00)	LA(99.59,99.64)	PH(99.45,95.00)
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.95	0.95	AT(98.47,95.00)	CH(98.74,95.00)	LA(99.23,98.82)	PH(98.97,95.00)
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.95	0.95	AT(98.40,96.72)	CH(98.02,95.00)	LA(98.77,98.03)	NY(99.98,99.84)
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.95	–	AT(97.29, –)	CH(95.00, –)	LA(98.16, –)	NY(99.94, -)
	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.98	AT(–, 99.62)	CH(–, 98.00)	LA(–, 99.86)	NY(–, 98.00)
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.98	0.98	AT(99.56,99.50)	CH(99.47,98.00)	LA(99.59,99.64)	PH(100.00,99.99)
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.98	0.98	AT(99.06,98.00)	CH(99.15,98.00)	LA(99.28,98.96)	NY(99.99,99.96)
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.98	0.98	AT(98.79,98.00)	CH(98.85,98.00)	LA(98.77,98.00)	NY(99.96,99.44)
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.98	–	CH(98.00, –)	LA(98.00, –)	MM(98.00, –)	PH(99.81, -)
0.75	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	NY(–, 90.00)	
	0.25	0.75	CH(99.19,80.66)	LA(99.58,99.60)	NY(99.10,66.09)	0.9	0.9	AT(99.55,99.52)	CH(99.39,96.13)	LA(99.59,99.62)	NY(99.32,90.00)
	0.5	0.5	CH(97.33,71.97)	LA(99.30,99.08)	NY(96.70,56.33)	0.9	0.9	AT(99.18,98.49)	CH(98.33,90.83)	LA(99.32,99.14)	NY(98.37,90.00)
	0.75	0.25	CH(91.22,62.14)	LA(98.83,98.20)	NY(87.96,46.37)	0.9	0.9	AT(97.27,92.48)	CH(96.94,90.00)	LA(98.78,98.05)	NY(97.00,90.00)
	1	0	CH(71.18, –)	LA(98.04, –)	NY(56.07, –)	0.9	–	AT(96.84, –)	CH(90.00, –)	LA(98.06, –)	NY(90.00, -)
	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.95	AT(–, 99.82)	CH(–, 98.23)	LA(–, 99.87)	NY(–, 95.00)
	0.25	0.75	CH(99.19,80.66)	LA(99.58,99.60)	NY(99.10,66.09)	0.95	0.95	AT(99.54,99.29)	CH(99.36,95.00)	LA(99.59,99.60)	NY(99.38,94.99)
	0.5	0.5	CH(97.33,71.97)	LA(99.30,99.08)	NY(96.70,56.33)	0.95	0.95	AT(98.58,95.00)	CH(98.76,94.99)	LA(99.24,98.84)	NY(98.86,95.00)
	0.75	0.25	CH(91.22,62.14)	LA(98.83,98.20)	NY(87.96,46.37)	0.95	0.95	AT(98.68,97.75)	CH(98.09,95.00)	LA(98.83,98.19)	NY(99.97,99.75)
	1	0	CH(71.18, –)	LA(98.04, –)	NY(56.07, –)	0.95	–	BA(99.38, –)	LA(98.03, –)	SL(95.00, –)	
	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.98	AT(–, 99.63)	CH(–, 98.00)	LA(–, 99.86)	NY(–, 98.00)
	0.25	0.75	CH(99.19,80.66)	LA(99.58,99.60)	NY(99.10,66.09)	0.98	0.98	AT(99.57,99.55)	CH(99.49,98.00)	LA(99.58,99.60)	NY(100.00,99.99)
	0.5	0.5	CH(97.33,71.97)	LA(99.30,99.08)	NY(96.70,56.33)	0.98	0.98	AT(99.18,98.63)	CH(99.18,97.99)	LA(99.28,99.00)	NY(99.99,99.94)
	0.75	0.25	CH(91.22,62.14)	LA(98.83,98.20)	NY(87.96,46.37)	0.98	0.98	BA(99.93,98.81)	LA(98.89,98.38)	SL(99.99,99.96)	
	1	0	CH(71.18, –)	LA(98.04, –)	NY(56.07, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)	

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6.

Table 5
Service levels at the hubs: hub node location model for $|N| = 25$, $a = 0.75$.

Parameters			Without service level constraints					With service level constraints						
α	f^e	f^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			β^e	β^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)						
0.5	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.9	AT(–, 99.82)	CH(–, 99.04)	LA(–, 99.87)	NY(–, 90.55)			
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.9	0.9	AT(99.57,99.49)	CH(99.38,96.41)	LA(99.59,99.64)	PH(99.31,90.00)			
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.9	0.9	AT(99.19,98.59)	CH(98.21,90.00)	LA(99.33,99.17)	PH(98.44,89.99)			
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.9	0.9	AT(97.27,92.57)	CH(96.89,89.99)	LA(98.76,98.02)	NY(97.08,90.00)			
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.9	–	AT(96.65, –)	CH(90.00, –)	LA(98.16, –)	NY(90.00, –)			
	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.95	AT(–, 99.81)	CH(–, 98.30)	LA(–, 99.87)	NY(–, 95.00)			
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.95	0.95	AT(99.49,99.17)	CH(99.34,95.00)	LA(99.59,99.64)	PH(99.45,95.00)			
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.95	0.95	AT(98.47,95.00)	CH(98.74,95.00)	LA(99.23,98.82)	PH(98.97,95.00)			
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.95	0.95	AT(98.65,97.65)	CH(98.06,95.00)	LA(98.78,98.08)	NY(97.99,95.00)	WA(98.82,98.20)		
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.95	–	CH(95.00, –)	DF(99.09, –)	LA(98.45, –)	NY(95.21, –)	WA(95.53, –)		
	0	1	CH(–, 93.05)	LA(–, 99.87)	PH(–, 50.03)	–	0.98	AT(–, 99.62)	CH(–, 98.00)	LA(–, 99.86)	NY(–, 98.00)			
	0.25	0.75	CH(99.26,88.03)	LA(99.59,99.63)	PH(99.00,42.02)	0.98	0.98	AT(99.44,98.00)	CH(99.45,98.00)	LA(99.45,98.95)	PH(99.55,98.00)			
	0.5	0.5	CH(97.76,80.96)	LA(99.32,99.15)	PH(95.96,33.89)	0.98	0.98	AT(99.26,98.92)	CH(99.13,98.00)	LA(99.28,99.03)	NY(99.15,98.00)	PG(99.05,98.00)		
	0.75	0.25	CH(93.28,72.32)	LA(98.88,98.34)	PH(83.67,26.46)	0.98	0.98	AT(98.72,98.00)	CH(98.80,98.00)	DF(99.20,99.28)	LA(98.95,98.56)	NY(98.82,98.00)	WA(98.77,98.00)	
	1	0	CH(79.81, –)	LA(98.14, –)	PH(34.05, –)	0.98	–	CH(98.00, –)	LA(98.00, –)	MM(98.00, –)	PH(99.81, –)			
0.75	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	NY(–, 90.00)				
	0.25	0.75	CH(99.19,80.66)	LA(99.58,99.60)	NY(99.10,66.09)	0.9	0.9	AT(99.55,99.52)	CH(99.39,96.13)	LA(99.59,99.62)	NY(99.32,90.00)			
	0.5	0.5	CH(97.33,71.97)	LA(99.30,99.08)	NY(96.70,56.33)	0.9	0.9	AT(99.18,98.49)	CH(98.33,90.83)	LA(99.32,99.14)	NY(98.37,90.00)			
	0.75	0.25	CH(91.22,62.14)	LA(98.83,98.20)	NY(87.96,46.37)	0.9	0.9	AT(97.27,92.48)	CH(96.94,90.00)	LA(98.78,98.05)	NY(97.00,90.00)			
	1	0	CH(71.18, –)	LA(98.04, –)	NY(56.07, –)	0.9	–	AT(96.84, –)	CH(90.00, –)	LA(98.06, –)	NY(90.00, –)			
	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.95	AT(–, 99.82)	CH(–, 98.23)	LA(–, 99.87)	NY(–, 95.00)			
	0.25	0.75	CH(99.19,80.66)	LA(99.58,99.60)	NY(99.10,66.09)	0.95	0.95	AT(99.54,99.29)	CH(99.36,95.00)	LA(99.59,99.60)	NY(99.38,94.99)			
	0.5	0.5	CH(97.33,71.97)	LA(99.30,99.08)	NY(96.70,56.33)	0.95	0.95	AT(98.58,95.00)	CH(98.76,94.99)	LA(99.24,98.84)	NY(98.86,95.00)			
	0.75	0.25	CH(91.22,62.14)	LA(98.83,98.20)	NY(87.96,46.37)	0.95	0.95	AT(98.68,97.75)	CH(98.09,95.00)	LA(98.83,98.19)	NY(99.97,99.75)			
	1	0	CH(71.18, –)	LA(98.04, –)	NY(56.07, –)	0.95	–	BA(99.38, –)	LA(98.03, –)	SL(95.00, –)				
	0	1	CH(–, 87.43)	LA(–, 99.86)	NY(–, 74.62)	–	0.98	AT(–, 99.63)	CH(–, 98.00)	LA(–, 99.86)	NY(–, 98.00)			
	0.25	0.75	CH(99.19,80.66)	,LA(99.58,99.60)	NY(99.10,66.09)	0.98	0.98	AT(99.37,97.99)	CH(99.49,98.00)	LA(99.51,98.99)	NY(99.52,98.00)			
	0.5	0.5	CH(97.33,71.97)	,LA(99.30,99.08)	NY(96.70,56.33)	0.98	0.98	AT(99.18,98.63)	CH(99.18,97.99)	LA(99.28,99.00)	NY(99.99,99.94)			
	0.75	0.25	CH(91.22,62.14)	,LA(98.83,98.20)	NY(87.96,46.37)	0.98	0.98	BA(99.95,99.34)	CH(98.78,98.00)	LA(98.89,98.38)	SL(98.78,98.05)			
	1	0	CH(71.18, –)	,LA(98.04, –)	NY(56.07, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)				

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6.

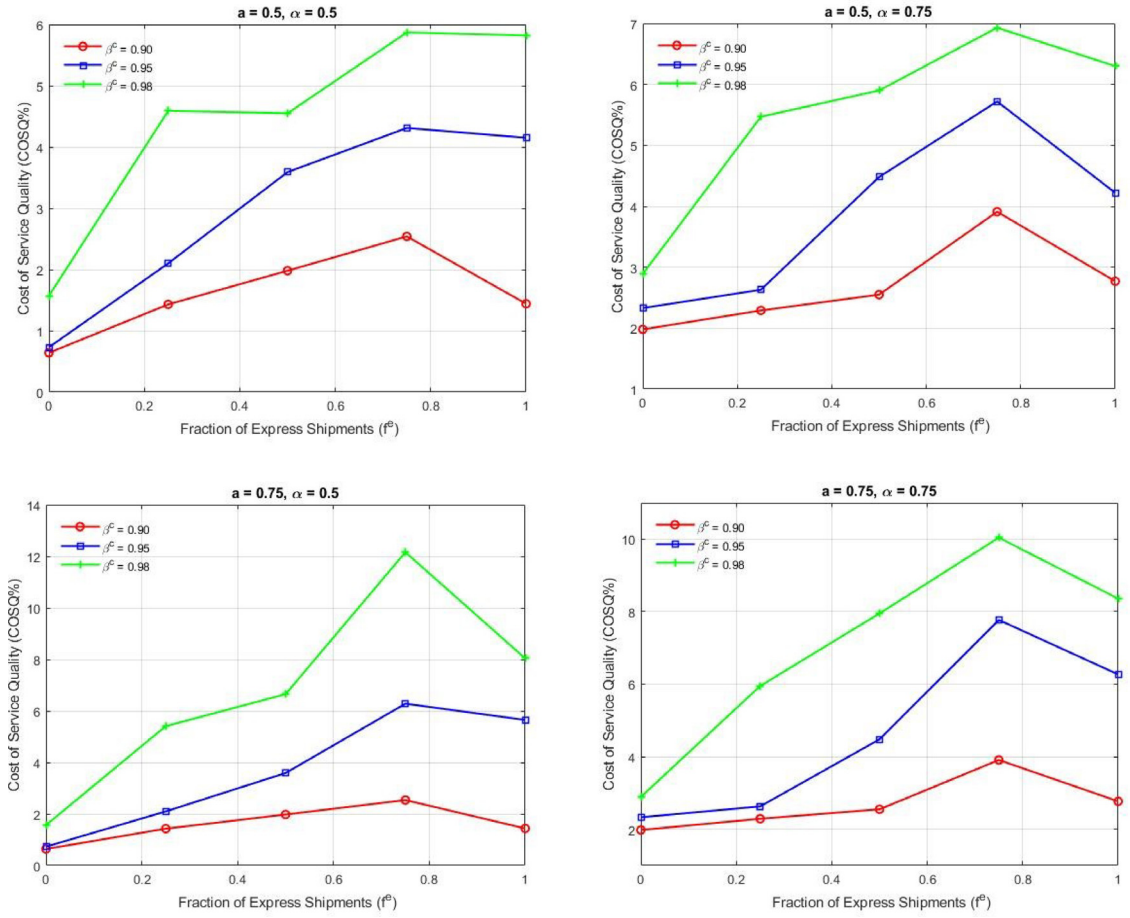

Fig. 3. Cost of service quality (CoSQ) vs. fraction of the express shipments (f^e) for hub node location model.

"Without Service Level Constraints" and "With Service Level Constraints". Results are presented for ($\beta^e = 90\%, \beta^r = 90\%$), ($\beta^e = 95\%, \beta^r = 95\%$), and ($\beta^e = 98\%, \beta^r = 98\%$) in case of "With Service Level Constraints". In these tables, $(f^e, f^r) = (0, 1)$ corresponds to the case with only one shipment class, for which the threshold on the maximum dwell time at a hub is $\tau^r = 10.0$ units. Similarly, $(f^e, f^r) = (1, 0)$ corresponds to the case with only one shipment class, for which the threshold on the maximum dwell time that at a hub is $\tau^e = 6.0$ units. Tables 2 and 3 report the resulting locations of hubs (abbreviations of cities) and their associated capacity levels, along with the total cost, number of iterations (Iter.) and time (CPU) taken by the cutting plane algorithm (Algorithm 1). Tables 4 and 5 report the achieved service levels (S^e_k, S^r_k) at different hubs in the resulting hub network for both 'Without Service Level Constraints" and "With Service Level Constraints".

From Tables 2, 3, we first note that without the service level requirements, the problems solve very quickly, requiring a maximum of less than 18 seconds. However, the maximum time required to solve the same problems with the service level constraints increases to over 800 seconds since the cutting plane algorithm often requires a couple of iterations to satisfy the service level requirements. Tables 4 and 5 clearly demonstrate that in the absence of any explicit service level requirements, the resulting hub network may provide very poor service levels. For example, for $a = 0.50$ or $a = 0.75$, $\alpha = 0.5$, $f^e = 1.0, f^r = 0.0$, the service level provided by the hub located at Philadelphia for the express shipments is as low as 34.05%. Tables 4 and 5 further show, as expected, that the service levels provided to the express and the regular shipments at the hubs in the network obtained without the service level requirements deteriorate with an increasing proportion of the express shipments in the system. They also show that increasing the discount (decreasing the value of α) on the inter-hub flows may sometimes result in the opening of more hubs in the presence of service level requirements to exploit the discounts arising from the consolidation of shipments at the hubs. The most interesting observation from these tables is highlighted in the following observation.

Observation 1. The configuration of the hub network with service level constraints on two shipment classes differs significantly from the one without these constraints.

4.2.2. Cost of service quality (CoSQ)

To measure the cost of providing a given level of service at hubs, we use the Cost of Service Quality (CoSQ%), which is the additional cost of the network design to guarantee a target service level (β^e, β^r) to both the shipment classes. It is computed as the difference between the total cost of the network design with and without the service level constraints, expressed as a percentage of the latter. CoSQ% for different parameter combinations are reported in Tables 2 and 3 for the Hub Node Location Model, and in Tables 6 and 7 (see Appendix C) for the Hub Arc Location Model. CoSQ% for different parameter combinations are also graphically presented in Figs. 3 and 4 for the Hub Node Location model, and the Hub Arc Location model, respectively. Figures 3 and 4 show, as expected, that CoSQ% is non-decreasing in β^c. However, the change in CoSQ% with an increase in the fraction of the express shipments (f^e) is not necessarily monotonic. Intuitively, an increase in the fraction of express shipments (f^e), that require a more stringent service level (i.e., a lower value of the maximum threshold

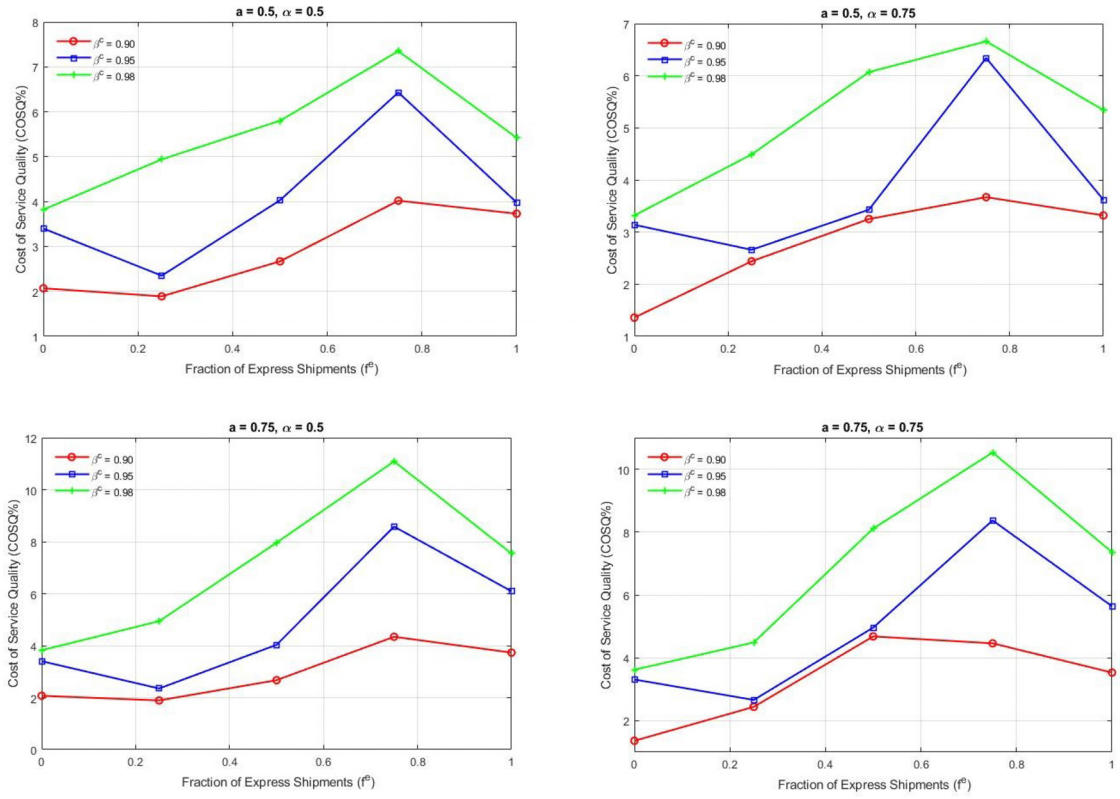

Fig. 4. Cost of service quality (CoSQ) vs. fraction of the express shipments (f^e) for hub arc location model with $q = 2$.

on dwell time), should increase the capacity, and hence the cost, required to meet their target service level. However, Figs. 3 and 4 show that the cost may decrease with an increase in the fraction of express shipments (e.g., when f^e increases from 0.75 to 1.0). This is an interesting observation, which is highlighted below as follows:

Observation 2. Extending priority service to a larger proportion of the customer base does not necessarily come at a cost; on the contrary, it may even reduce cost.

The above seemingly counter-intuitive observation can be explained as follows. An increase in f^e is also accompanied by a corresponding decrease in the fraction of the regular shipments (since $f^r = 1 - f^e$) that receive less preferential treatment at the hubs, thereby decreasing the capacity required to meet their service level requirement. Hence, in the presence of the priority in service, two opposite forces come into play: (i.) an increase in the capacity required to serve the additional express shipments; and (ii.) a decrease in the capacity required to serve fewer regular shipments. The net result may be either an increase, decrease, or no change in the capacity required, and hence a corresponding increase, decrease, or no change in CoSQ%. We illustrate this phenomenon using a simple example of a single-queue system.

Illustrative example: Consider a preemptive priority M/M/1 queue with two customer classes, which is used to model each of the hubs in the hub (node or arc) location model. Let the mean arrival rate of the customers be $\lambda = 1.0$, and the composition of the high-priority (express) and low-priority (regular) customers be given as $f^e = 0.25$ and $f^r = 0.75$; thus, their mean arrival rates being $\lambda^e = f^e \times \lambda = 0.25$ and $\lambda^r = f^r \times \lambda = 0.75$, respectively. Further, assume that the maximum threshold on the dwell for the high-priority customers is $\tau^e = 6.0$ unit, that for the low-priority customers is $\tau^r = 7.0$ unit, and their required service levels are $\beta^c(\tau^c) = P\{W^c \le \tau^c\} = 0.95\ \forall c \in \{e, r\}$. Let the cost of server capacity be \$1 per unit service rate.

For the special case in which all customers are high-priority, i.e., $f^e = 1.0$ (and $f^r = 0.0$), the service discipline reduces to FCFS, in which case the minimum required service capacity is given as $\mu = \lambda^e - \ln(1 - \beta^e)/\tau^e$ (from (32)). For $\tau^e = 6.0$ and $\beta^e = 0.95$, this gives $\mu = 5.993$, and hence the cost of service capacity = \$5.993. At the same mean arrival rate ($\lambda = 1.0$) and service capacity ($\mu = 5.993$), but for $f^e = 0.25$ and $f^r = 0.75$, the service level achieved for the high-priority customers can be computed as $S^e(\tau^e) = P\{W^e \le \tau^e\} = 1 - e^{-(\mu - \lambda^e)/\tau^e} = 0.95696$, which satisfies the service level requirement for the high-priority customers of $\beta^e = 0.95$. In the absence of any closed-form expression for the service level of the low-priority customers in a preemptive-priority M/M/1 queue, computing it numerically using the matrix geometric method, described in Section 3.1, gives $S^r(\tau^r) = P\{W^r \le \tau^r\} = 0.94431$, which is strictly less than the required service level of $\beta^r = 0.95$. Hence, to increase the service level of the low-priority customers to $\beta^r = 0.95$, the service capacity needs to be increased beyond $\mu = 5.993$, costing > \$5.993. This corroborates the above observation that providing a service level (of 95%) to a higher proportion ($f^e = 1$) of high-priority customers comes at a lower cost (\$5.993) than providing the same service level to a lower proportion ($f^e = 0.25$) of high-priority customers (which costs strictly greater than \$5.993).

5. Conclusions

In this paper, we studied the hub node and hub arc location problems, characterized by stochastic demand and congestion, with an explicit consideration for shipment heterogeneity. Shipments were thus assumed to belong to two different priority classes, express and regular, with the express customers always receiving priority in service at the hubs. To account for the heterogeneous ship-

ment requirements, we used a different service level constraint, defined as a lower limit on the probability of a shipment waiting for more than a given threshold at a hub, for each shipment class. The network of hubs, given their locations, was thus modeled as spatially distributed preemptive priority M/M/1 queues. The model sought to determine the hub-and-spoke network design at the minimum total cost, which included the total fixed cost of equipping the open hubs with sufficient processing capacity and the variable transportation costs, subject to a service level constraint for each shipment class. The problem proved to be challenging, especially in the absence of any known analytical expression for the sojourn time distribution of the regular shipments in a preemptive priority M/M/1 queue. To this end, we developed a solution technique that uses the matrix geometric method in a cutting plane framework. Based on our computational study, we demonstrated that the optimal network configuration that accounts for the different service levels demanded by heterogeneous customer classes may differ significantly from the one that does not consider the service level constraints. Further, we observed that increasing the fraction of shipments that receive priority in service may not necessarily increase the total cost of the network design.

The work presented in this paper can be extended in a number of ways. Our study is based on the assumption that each hub behaves like a preemptive priority M/M/1 queue. An immediate extension of the current work will be to consider a non-preemptive priority discipline at hubs. Another possible extension would be a more generalized queuing model, like a priority M/G/1 queue model, of the hubs, although the resulting model will be extremely challenging to solve.

Acknowledgments

The authors thank the editor and two anonymous reviewers for their valuable comments on a previous version of this paper. This research was supported by the Research & Publication Grant, Indian Institute of Management Ahmedabad to the first author, and by the National Science and Engineering Research Council of Canada (NSERC) grant to the second author.

Appendix A

$$A_0 = \begin{pmatrix} \Lambda_k^r & & & & \\ & \Lambda_k^r & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \Lambda_k^r \end{pmatrix};$$

$$A_2 = \begin{pmatrix} \mu_k & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix};$$

$$B_0 = \begin{pmatrix} * & \Lambda_k^e & & & \\ \mu_k & * & \Lambda_k^e & & \\ & \mu_k & * & \Lambda_k^e & \\ & & \ddots & \ddots & \ddots \\ & & & \mu_k & * \end{pmatrix}$$

where $*$ is such that $A_0\mathbf{1} + B_0\mathbf{1} = \mathbf{0}$. $A_1 = B_0 - A_2$. Here, $\mathbf{1}$ is a column vector of ones of size $M+1$.

Appendix B

From Section 3.1.2, we have:

$$\overline{(S_k^r(\tau^r))^p} = \sum_{n=0}^{\infty} d_n e^{-\gamma\tau^r} \frac{(\gamma\tau^r)^n}{n!} \tag{A1}$$

where, d_n is given by:
$$d_n = \sum_{i=0}^{\infty} \mathbf{x}_0 R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{1}, \quad n \geq 0 \tag{A2}$$

Now,

$$\begin{aligned}
&\sum_{i=0}^{\infty} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{1} \\
&= \sum_{i=0}^{n+1} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{1} + \sum_{i=n+2}^{\infty} R^i \sum_{v=0}^{n} G_v^{(n)}\mathbf{1} \qquad \left(\text{since} \quad G_v^{(n)} = 0 \quad \text{for} \quad v > n\right) \\
&= \sum_{v=0}^{n+1}\sum_{i=v}^{n+1} R^i G_v^{(n)}\mathbf{1} + (I-R)^{-1}R^{n+2}\mathbf{1} \qquad \left(\text{since} \quad \sum_{v=0}^{n} G_v^{(n)}\mathbf{1} = \mathbf{1}\right) \\
&= \sum_{v=0}^{n+1} (I-R)^{-1}(R^v - R^{n+2}) G_v^{(n)}\mathbf{1} + (I-R)^{-1}R^{n+2}\mathbf{1} \\
&= \sum_{v=0}^{n} (I-R)^{-1}R^v G_v^{(n)}\mathbf{1} + (I-R)^{-1}R^{n+1}G_{n+1}^{(n)}\mathbf{1} \qquad \left(\text{since} \quad \sum_{v=0}^{n+1} G_v^{(n)}\mathbf{1} = \mathbf{1}\right) \\
&= \sum_{v=0}^{n} (I-R)^{-1}R^v G_v^{(n)}\mathbf{1} \qquad \left(\text{since} \quad G_v^{(n)} = 0 \quad \text{for} \quad v > n\right) \\
&= (I-R)^{-1}H_n\mathbf{1} \qquad n \geq 0
\end{aligned} \tag{A3}$$

where, $H_n = \sum_{v=0}^{n} R^v G_v^{(n)}$. Substituting (A3) in (A2), and then substituting (A2) in (A1) gives:

$$(S_k^r(\tau^r))^p = 1 - \overline{(S_k^r(\tau^r))^p} = 1 - \sum_{n=0}^{\infty} e^{-\gamma\tau^r} \frac{(\gamma\tau^r)^n}{n!} \mathbf{x}_0 (I-R)^{-1} H_n \mathbf{1}$$

Appendix C

Table 6
Effect of service level constraints: hub arc location model for $|N| = 25$, $q = 2$, $a = 0.50$.

Parameters			Results without service level constraints						Results with service level constraints									
α	f^e	f^r	Hub (capacity level)			Cost	Iter.	CPU	β^e	β^r	Hub (capacity level)				Cost	Iter.	CPU	CoSQ%
0.5	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	206.08	–	0.9	CH(1)	LA(1)	NY(1)		2577.77	1	4611.22	2.07
	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	87.01	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2487.36	4	439.22	1.89
	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	112.18	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2535.34	5	963.37	2.67
	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	357.49	0.9	0.9	CH(1)	LA(1)	PH(2)		2597.81	5	3478.25	4.02
	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	190.18	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2619.68	1	2965.2	3.73
	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	228.86	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2611.5	1	5385.78	3.40
	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	88.04	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2498.61	5	603.38	2.35
	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	116.34	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2568.97	7	1888.77	4.03
	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	367.69	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2658.1	4	8073.19	6.43
	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	195.51	0.95	–	LA(1)	PH(2)	SL(1)		2625.95	1	1661.21	3.98
	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	211.25	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2621.98	1	3068.31	3.82
	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	88.53	0.98	0.98	AT(1)	CH(1)	LA(1)	PH(1)	2561.96	6	710.81	4.94
	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	115.24	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2612.49	5	3739.66	5.80
	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	364.7	0.98	0.98	BA(2)	LA(1)	SL(2)		2681.02	3	5993.53	7.35
	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	196.89	0.98	–	BA(2)	LA(1)	SL(1)		2662.56	1	2763.16	5.42
0.75	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	136.93	–	0.9	CH(1)	LA(1)	PH(1)		2685.15	1	1810.07	1.36
	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	73.69	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2683.01	5	458.7	2.44
	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	104.41	0.9	0.9	LA(1)	PH(2)	SL(1)		2715.84	4	290.66	3.25
	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	86.7	0.9	0.9	CH(1)	LA(1)	PH(2)		2736.61	5	525.34	3.67
	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	134.93	0.9	–	LA(1)	PH(2)	SL(1)		2736.96	1	1069.86	3.32
	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	141.73	–	0.95	LA(1)	PH(2)	SL(1)		2732.27	1	3225.72	3.14
	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	74.78	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2688.84	5	442.59	2.66
	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	108.32	0.95	0.95	LA(1)	PH(2)	SL(1)		2720.65	4	365.68	3.43
	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	83.57	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2807.11	4	3594.62	6.34
	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	139.47	0.95	–	LA(1)	PH(2)	SL(1)		2744.83	1	1001.88	3.62
	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	141.39	–	0.98	LA(1)	PH(2)	SL(1)		2736.99	1	1079.61	3.32
	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	74.78	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2736.75	6	432.79	4.49
	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	107.42	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2790.23	5	6217.09	6.07
	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	83.52	0.98	0.98	BA(2)	LA(1)	SL(2)		2815.48	3	1762.38	6.66
	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	137.92	0.98	–	BA(2)	LA(1)	SL(1)		2790.6	1	1968.58	5.34

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6; CPU = Computation Time (in seconds); CoSQ = Cost of Service Quality.

Table 7
Effect of service level constraints: hub arc location model for $|N| = 25$, $q = 2$, $a = 0.75$.

Parameters			Results without service level constraints						Results with service level constraints									
α	f^e	f^r	Hub (capacity level)			Cost	Iter.	CPU	β^e	β^r	Hub (capacity level)				Cost	Iter.	CPU	CoSQ%
0.5	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	298.4	–	0.9	CH(1)	LA(1)	NY(1)		2577.77	1	8272.57	2.07
0.5	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	60.05	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2487.36	4	619.98	1.89
0.5	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	106.88	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2535.34	5	2322.97	2.67
0.5	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	274.91	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2605.95	5	7776.66	4.34
0.5	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	296.16	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2619.68	1	2177.06	3.73
0.5	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	327.62	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2611.5	1	6087.22	3.40
0.5	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	61.2	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2498.61	5	832.05	2.35
0.5	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	107.29	0.95	0.95	AT(1)	CH(1)	LA(1)	PH(1)	2568.97	7	2145.25	4.03
0.5	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	288.66	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2711.66	5	10264.24	8.58
0.5	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	306.7	0.95	–	LA(1)	PH(2)	SL(1)		2679.47	1	8397.79	6.09
0.5	0	1	CH(1)	LA(1)	PH(1)	2525.55	1	313.21	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2621.98	1	3788.47	3.82
0.5	0.25	0.75	CH(1)	LA(1)	PH(1)	2441.25	1	61.03	0.98	0.98	AT(1)	CH(1)	LA(1)	PH(1)	2561.93	6	1012.99	4.94
0.5	0.5	0.5	CH(1)	LA(1)	PH(1)	2469.35	1	106.62	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2666.01	5	7072.01	7.96
0.5	0.75	0.25	CH(1)	LA(1)	PH(1)	2497.45	1	289.28	0.98	0.98	BA(2)	CH(1)	LA(1)	SL(1)	2774.41	6	17888.95	11.09
0.5	1	0	CH(1)	LA(1)	PH(1)	2525.55	1	305.28	0.98	–	BA(2)	LA(1)	SL(1)		2716.08	1	4609.52	7.54
0.75	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	85.97	–	0.9	CH(1)	LA(1)	PH(1)		2685.15	1	1965.56	1.36
0.75	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	77.64	0.9	0.9	AT(1)	CH(1)	LA(1)	NY(1)	2683.01	5	739.95	2.44
0.75	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	64.7	0.9	0.9	CL(1)	LA(1)	PH(1)	SL(1)	2753.5	5	894.22	4.68
0.75	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	87.04	0.9	0.9	AT(1)	CH(1)	LA(1)	PH(1)	2757.49	5	1568.18	4.46
0.75	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	82.9	0.9	–	AT(1)	CH(1)	LA(1)	NY(1)	2742.64	1	1443.52	3.53
0.75	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	90.76	–	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2736.6	1	2593.15	3.31
0.75	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	79.11	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(1)	2688.84	5	606.25	2.66
0.75	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	66.22	0.95	0.95	CL(1)	LA(1)	PH(1)	SL(1)	2760.87	6	979.19	4.96
0.75	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	86.01	0.95	0.95	AT(1)	CH(1)	LA(1)	NY(2)	2860.62	5	3861.5	8.37
0.75	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	86.64	0.95	–	LA(1)	PH(2)	SL(1)		2798.35	1	1780.88	5.64
0.75	0	1	LA(1)	PH(1)	SL(1)	2649.02	1	89.71	–	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2744.84	1	1323.51	3.62
0.75	0.25	0.75	CH(1)	LA(1)	NY(1)	2619.09	1	78.43	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(1)	2736.75	6	530.79	4.49
0.75	0.5	0.5	LA(1)	PH(1)	SL(1)	2630.45	1	66.4	0.98	0.98	AT(1)	CH(1)	LA(1)	NY(2)	2843.74	5	5751.36	8.11
0.75	0.75	0.25	LA(1)	PH(1)	SL(1)	2639.73	1	85.53	0.98	0.98	BA(2)	CH(1)	LA(1)	SL(1)	2917.61	6	6395.85	10.53
0.75	1	0	LA(1)	PH(1)	SL(1)	2649.02	1	86.19	0.98	–	BA(2)	LA(1)	SL(1)		2844.12	1	3396.1	7.36

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6; CPU = Computation Time (in seconds); CoSQ = Cost of Service Quality.

Table 8
Service levels at the hubs: hub arc location model for $|N| = 25$, $q = 2$, $a = 0.50$.

Parameters			Without service level constraints					With service level constraints			
α	f^e	f^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			β^e	β^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	NY(–, 90.00)	
0.5	0.25	0.75	CH(99.11,83.94)	LA(99.59,99.63)	PH(99.16,55.05)	0.9	0.9	AT(99.59,99.57)	CH(99.24,95.68)	LA(99.59,99.63)	PH(99.41,89.99)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.9	0.9	AT(99.18,98.57)	CH(97.93,90.00)	LA(99.32,99.16)	PH(98.67,90.00)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.9	0.9	CH(97.25,90.00)	LA(98.66,97.71)	PH(99.92,98.62)	
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.9	–	AT(97.04, –)	CH(90.00, –)	LA(97.93, –)	NY(90.00, -)
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.95	AT(–, 99.91)	CH(–, 95.47)	LA(–, 99.87)	NY(–, 96.11)
0.5	0.25	0.75	CH(99.11,83.94)	LA(99.59,99.63)	PH(99.16,55.05)	0.95	0.95	AT(99.50,99.18)	CH(99.28,95.00)	LA(99.59,99.63)	PH(99.49,94.99)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.95	0.95	AT(98.52,95.00)	CH(98.58,95.00)	LA(99.23,98.80)	PH(99.06,95.00)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.95	0.95	AT(98.72,97.76)	CH(98.02,94.99)	LA(98.79,98.10)	NY(99.98,99.76)
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.95	–	LA(97.46, –)	PH(99.52, –)	SL(95.00, –)	
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.98	AT(–, 99.67)	CH(–, 98.00)	LA(–, 99.84)	NY(–, 98.00)
0.5	0.25	0.75	CH(99.11,83.94)	LA(99.59,99.63)	PH(99.16,55.05)	0.98	0.98	AT(99.41,98.00)	CH(99.48,98.00)	LA(99.44,98.93)	PH(99.56,98.00)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.98	0.98	AT(99.30,98.99)	CH(99.21,98.00)	LA(99.24,98.91)	NY(99.99,99.92)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.98	0.98	BA(99.95,99.24)	LA(98.90,98.40)	SL(99.99,99.91)	
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)	
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	PH(–, 90.00)	
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.9	0.9	AT(99.59,99.64)	CH(99.28,95.28)	LA(99.59,99.61)	NY(99.38,90.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.89,20.66)	SL(97.79,83.76)	0.9	0.9	LA(99.24,98.90)	PH(99.99,99.77)	SL(98.45,90.09)	
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.9	0.9	CH(97.11,90.00)	LA(98.60,97.50)	PH(99.92,98.72)	
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.9	–	LA(97.66, –)	PH(99.73, –)	SL(90.26, –)	
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.95	LA(–, 99.87)	PH(–, 100.00)	SL(–, 95.00)	
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.95	0.95	AT(99.54,99.31)	CH(99.29,95.00)	LA(99.58,99.59)	NY(99.44,95.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.89,20.66)	SL(97.79,83.76)	0.95	0.95	LA(99.22,98.83)	PH(99.98,99.58)	SL(98.86,94.99)	
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.95	0.95	AT(98.99,98.67)	CH(98.12,94.99)	LA(98.64,97.65)	NY(99.97,99.69)
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.95	–	LA(97.49, –)	PH(99.51, –)	SL(95.00, –)	
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.98	LA(–, 99.81)	PH(–, 99.99)	SL(–, 98.00)	
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.98	0.98	AT(99.42,98.00)	CH(99.49,98.00)	LA(99.45,98.95)	NY(99.53,98.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.89,20.66)	SL(97.79,83.76)	0.98	0.98	AT(99.29,99.04)	CH(99.23,98.00)	LA(99.23,98.84)	NY(99.99,99.92)
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.98	0.98	BA(99.95,99.21)	LA(98.89,98.38)	SL(99.99,99.92)	
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)	

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6.

Table 9
Service levels at the hubs: hub arc location model for $|N| = 25$, $q = 2$, $a = 0.75$.

Parameters			Without service level constraints					With service level constraints			
α	f^e	f^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			β^e	β^r	Hub($S_k^e(\tau^e = 6), S_k^r(\tau^r = 10)$)			
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	NY(–, 90.00)	
0.5	0.25	0.75	CH(99.09,83.26)	LA(99.59,99.63)	PH(99.18,56.75)	0.9	0.9	AT(99.59,99.57)	CH(99.24,95.68)	LA(99.59,99.63)	PH(99.41,89.99)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.9	0.9	AT(99.18,98.57)	CH(97.93,90.00)	LA(99.32,99.16)	PH(98.67,90.00)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.9	0.9	AT(97.22,92.68)	CH(96.65,90.00)	LA(98.76,97.96)	PH(97.35,90.00)
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.9	–	AT(97.04, –)	CH(90.00, –)	LA(97.93, –)	NY(90.00, -)
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.95	AT(–, 99.91)	CH(–, 96.48)	LA(–, 99.87)	NY(–, 95.00)
0.5	0.25	0.75	CH(99.09,83.26)	LA(99.59,99.63)	PH(99.18,56.75)	0.95	0.95	AT(99.50,99.18)	CH(99.28,95.00)	LA(99.59,99.63)	PH(99.49,94.99)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.95	0.95	AT(98.52,95.00)	CH(98.58,95.00)	LA(99.23,98.80)	PH(99.06,95.00)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.95	0.95	AT(98.72,97.76)	CH(98.02,95.00)	LA(98.79,98.10)	NY(99.98,99.76)
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.95	–	LA(97.46, –)	PH(99.52, –)	SL(95.00, –)	
0.5	0	1	CH(–, 75.12)	LA(–, 99.87)	PH(–, 86.05)	–	0.98	AT(–, 99.67)	CH(–, 98.00)	LA(–, 99.84)	NY(–, 98.00)
0.5	0.25	0.75	CH(99.09,83.26)	LA(99.59,99.63)	PH(99.18,56.75)	0.98	0.98	AT(99.41,98.00)	CH(99.48,98.00)	LA(99.44,98.93)	PH(99.56,98.00)
0.5	0.5	0.5	CH(96.80,70.53)	LA(99.32,99.15)	PH(97.17,55.39)	0.98	0.98	AT(99.30,98.99)	CH(99.21,98.00)	LA(99.24,98.91)	NY(99.99,99.92)
0.5	0.75	0.25	CH(88.07,53.37)	LA(98.88,98.34)	PH(90.80,54.34)	0.98	0.98	BA(99.96,99.37)	CH(98.75,98.00)	LA(98.86,98.31)	SL(98.80,98.00)
0.5	1	0	CH(56.59, –)	LA(98.14, –)	PH(69.32, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)	
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.9	CH(–, 90.00)	LA(–, 99.55)	PH(–, 90.00)	
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.9	0.9	AT(99.59,99.64)	CH(99.28,95.28)	LA(99.59,99.61)	NY(99.38,90.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.91,21.11)	SL(97.77,83.64)	0.9	0.9	CL(98.89,97.18)	LA(99.33,99.17)	PH(98.60,90.00)	SL(98.54,94.19)
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.9	0.9	AT(97.30,92.83)	CH(96.74,90.00)	LA(98.73,97.91)	PH(97.27,90.00)
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.9	–	AT(97.02, –)	CH(90.00, –)	LA(97.94, –)	NY(90.00, -)
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.95	AT(–, 99.91)	CH(–, 96.15)	LA(–, 99.86)	NY(–, 95.90)
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.95	0.95	AT(99.54,99.31)	CH(99.29,95.00)	LA(99.58,99.59)	NY(99.44,95.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.91,21.11)	SL(97.77,83.64)	0.95	0.95	CL(98.64,94.99)	LA(99.22,98.82)	PH(98.89,95.00)	SL(98.71,95.00)
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.95	0.95	AT(98.99,98.67)	CH(98.12,95.00)	LA(98.64,97.65)	NY(99.97,99.69)
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.95	–	LA(97.49, –)	PH(99.51, –)	SL(95.00, –)	
0.75	0	1	LA(–, 99.87)	PH(–, 48.36)	SL(–, 93.11)	–	0.98	AT(–, 99.66)	CH(–, 98.00)	LA(–, 99.85)	NY(–, 98.00)
0.75	0.25	0.75	CH(99.18,80.37)	LA(99.58,99.60)	NY(99.10,66.56)	0.98	0.98	AT(99.42,98.00)	CH(99.49,98.00)	LA(99.45,98.95)	NY(99.53,98.00)
0.75	0.5	0.5	LA(99.33,99.17)	PH(95.91,21.11)	SL(97.77,83.64)	0.98	0.98	AT(99.34,99.19)	CH(99.23,98.00)	LA(99.22,98.82)	NY(99.99,99.91)
0.75	0.75	0.25	LA(98.89,98.38)	PH(83.43,21.02)	SL(93.31,73.88)	0.98	0.98	BA(99.96,99.45)	CH(98.78,98.00)	LA(98.75,98.00)	SL(98.76,98.00)
0.75	1	0	LA(98.16, –)	PH(32.73, –)	SL(79.92, –)	0.98	–	BA(98.46, –)	LA(98.00, –)	SL(98.00, –)	

$(f^e, f^r) = (0, 1)$ refers to a single shipment class with the maximum threshold on sojourn time = 10, whereas $f^e, f^r = (1, 0)$ refers to a single shipment class with the maximum threshold on sojourn time = 6.

References

Abate, J., & Whitt, W. (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems, 25*(1–4), 173–233.

Abdinnour-Helm, S., & Venkataramanan, M. A. (1998). Solution approaches to hub location problems. *Annals of Operations Research, 78*, 31–50.

Alkaabneh, F., Diabat, A., & Elhedhli, S. (2019). A lagrangian heuristic and grasp for the hub-and-spoke network system with economies-of-scale and congestion. *Transportation Research Part C: Emerging Technologies, 102*, 249–273.

Alumur, S., & Kara, B. (2008). Network hub location problems: The state-of-the-art. *European Journal of Operational Research, 190*, 1–21.

Alumur, S., Kara, B., & Karasan, O. (2009). The design of single allocation incomplete hub-networks. *Transportation Research B, 43*, 936–951.

Alumur, S. A., Campbell, J. F., Contreras, I., Kara, B. Y., Marianov, V., & O'Kelly, M. E. (2021). Perspectives on modeling hub location problems. *European Journal of Operational Research, 291*, 1–17.

Alumur, S. A., Nickel, S., & Saldanha-da Gama, F. (2012). Hub location under uncertainty. *Transportation Research Part B: Methodological, 46*(4), 529–543.

Atlason, J., Epelman, M., & Henderson, S. (2004). Call center staffing with simulation and cutting plane methods. *Annals of Operations Research, 127*(1–4), 333–358.

Aykin, T. (1994). Lagrangean relaxation based appraoches to capacitated hub-and-spoke network design problem. *European Journal of Operational Research, 79*(3), 501–523.

Azizi, N., Vidyathi, N., & Chauhan, S. S. (2018). Modelling and analysis of hub-and-spoke networks under stochastic demand and congestion. *Annals of Operations Research, 264*, 1–40.

Bhatt, S. D., Jayaswal, S., Sinha, A., & Vidyarthi, N. (2021). Alternate second order conic program reformulations for hub location under stochastic demand and congestion. *Annals of Operations Research, 304*(1), 481–527.

Boland, N., Krishnamoorthy, M., Ernst, A. T., & Ebery, J. (2004). Preprocessing and cutting for multiple allocation hub location problems. *European Journal of Operational Research, 155*(3), 638–653.

Bryan, D. (1998). Extensions to the hub location problem: Formulations and numerical examples. *Geographical Analysis, 30*, 315–330.

Camargo, R., de Miranda, G., Jr., & Ferreira, R. (2011a). A hybrid outer-approximation/benders decomposition algorithm for the single allocation hub location problem under congestion. *Operations Research Letters, 39*(12), 329–337.

Camargo, R., de Miranda, G., Jr., Ferreira, R., & Luna, H. (2009). Multiple allocation hub-and-spoke network design under hub congestion. *Computers and Operations Research, 36*(12), 3097–3106.

Camargo, R. S., de Miranda, G., Jr., & Ferreira, R. P. (2011b). A hybrid outer-approximation/benders decomposition algorithm for the single allocation hub location problem under congestion. *Operations Research Letters, 39*(5), 329–337.

Campbell, J. (1994b). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research, 72*, 387–405.

Campbell, J. (2009). Hub location for time definite transportation. *Computers and Operations Research, 36*, 3107–3116.

Campbell, J. (2013). Modeling economies of scale in transporation hub networks. In *Proceedings of the 46th Hawaii international conference on systems sciences* (pp. 1154–1163). http://origin-www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892b154.pdf.

Campbell, J., Ernst, A., & Krishnamoorthy, M. (2002). Hub location problems. In Z. Drezner, & H. Hamacher (Eds.), *Location analysis: Theory and applications* (pp. 373–408). Springer, Berlin.

Campbell, J., Ernst, A., & Krishnamoorthy, M. (2005a). Hub arc location problems: Part I—Introduction and results. *Management Science, 51*, 1540–1555.

Campbell, J., Ernst, A., & Krishnamoorthy, M. (2005b). Hub arc location problems: Part II—Formulations and optimal algorithms. *Management Science, 51*, 1556–1571.

Campbell, J., & O'Kelly, M. (2012). Twenty-five years of hub location research. *Transportation Science, 46*(2), 153–169.

Canovas, L., Garcia, S., & Marin, A. (2007). Solving the uncapacitated multiple allocation hub location problem by means of a dual-ascent technique. *European Journal of Operations Research, 179*, 990–1007.

Contreras, I. (2021). Hub network design. In T. G. Crainic, M. Gendreau, & B. Gendron (Eds.), *Network design with applications to transportation and logistics* (pp. 567–598). Springer International Publishing.

Contreras, I., Cordeau, J.-F., & Laporte, G. (2012). Benders decomposition for large-scale uncapacitated hub location problem. *Operations Research, 59*(6), 1477–1490.

Contreras, I., & O'Kelly, M. (2019). Hub location problems. In T. G. Crainic, M. Gendreau, & B. Gendron (Eds.), *Location science* (pp. 327–357). Springer International Publishing.

Correia, I., Nickel, S., & Saldanha-da-Gama, F. (2010). Single-assignment hub location problems with multiple capacity levels. *Transportation Research: Part B, 44*, 1047–1066.

Costa, M., Captivo, M. E., & Climaco, J. (2008). Capacitated single allocation hub location problem—A bi-criteria approach. *Computers and Operations Research, 35*, 3671–3695.

Cunha, C., & Silva, M. (2007). A genetic algorithm for the problem of configuring a hub-and-spoke network for a LTL trucking company in Brazil. *European Journal of Operational Research, 179*, 747–758.

Ebery, J., Krishnamoorthy, M., Ernst, A., & Boland, N. (2000). The capacitated multiple allocation hub location problem: Formulations and algorithms. *European Journal of Operational Research, 120*, 614–631.

Elhedhli, S., & Hu, F. X. (2005). Hub-and-spoke network design with congestion. *Computers and Operations Research, 32*, 1615–1632.

Elhedhli, S., & Wu, H. (2010). A Lagrangean heuristic for hub-and-spoke system design with capacity selection and congestion. *INFORMS Journal of Computing, 22*(2), 282–296.

Ernst, A., & Krishnamoorthy, M. (1999). Solution algorithms for the capacitated single allocation hub location problem. *Annals of Operations Research, 86*, 141–159.

Gross, D., & Harris, C. (1998). *Fundamentals of queueing theory* (3rd ed.). New York: John Wiley and Sons.

Guldmann, J., & Shen, G. (1997). A general mixed integer nonlinear optimization model for hub network design. *Working paper*. Columbus, Ohio: Department of City and Regional Planning, The Ohio State University.

Hamacher, H. W., Labbé, M., Nickel, S., & Sonneborn, T. (2004). Adapting polyhedral properties from facility to hub location problems. *Discrete Applied Mathematics, 145*(1), 104–116.

Hasanzadeh, H., Bashiri, M., & Amiri, A. (2018). A new approach to optimize a hub covering location problem with a queue estimation component using genetic programming. *Soft Computing, 22*(3), 949–961.

Ishfaq, R., & Sox, C. R. (2012). Design of intermodal logistics networks with hub delays. *European Journal of Operational Research, 220*(3), 629–641.

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

Jayaswal, S., & Vidyarthi, N. (2017). Facility location under service level constraints for heterogeneous customers. *Annals of Operations Research, 253*(1), 275–305.

Kian, R., & Kargar, K. (2016). Comparison of the formulations for a hub-and-spoke network design problem under congestion. *Computers and Industrial Engineering, 101*, 504–512.

Kimms, A. (2006). Economies of scale in hub and spoke network design models: We have it all wrong. In M. Morlock, C. Schwindt, N. Trautmann, & J. Zimmermann (Eds.), *Perspective on operations research* (pp. 293–317). Wiesbaden: Deutscher-Universitats Verlag/GWV Fachverlage GmbH.

Labbe, M., & Yaman, H. (2004). Projecting the flow variables for hub location problems. *Network, 44*(2), 84–93.

Labbe, M., Yaman, H., & Gourdin, E. (2005). A branch and cut algorithm for hub location problems with single assignment. *Mathematical Programming, 102*, 371–405.

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Philadelphia, USA: Society for Industrial and Applied Mathematics.

Marianov, V., & Serra, D. (2003). Location models for airline hubs behaving as M/D/c queues. *Computer and Operations Research, 30*, 983–1003.

Marin, A. (2005a). Formulating and solving splittable capacitated multiple allocation hub location problems. *Computer and Operations Research, 32*, 3093–3109.

Marin, A. (2005b). Uncapacitated Euclidean hub location: Strengthened formulation, new facets and a relax-and-cut algorithm. *Journal of Global Optimization, 33*, 393–422.

Marıń, A., Cánovas, L., & Landete, M. (2006). New formulations for the uncapacitated multiple allocation hub location problem. *European Journal of Operational Research, 172*(1), 274–292.

Miller, D. (1981). Computation of steady-state probabilities for M/M/1 priority queues. *Operations Research, 29*(5), 945–958.

Mohammadi, M., Jolai, F., & Rostami, H. (2011). An M/M/c queue model for hub covering location problem. *Mathematical and Computer Modelling, 54*(11–12), 2623–2638.

Neuts, M. (1981). *Matrix-Geometric solutions in stochastic models: An algorithmic approach*. Mineola, USA: Dover Publications.

O'Kelly, M. (1986a). The location of interacting hub facilities. *Transportation Science, 20*, 92–106.

O'Kelly, M. (1992b). Hub facility location with fixed costs. *Papers in Regional Science, 71*(3), 293–306.

O'Kelly, M., & Bryan, D. (1998). Hub location with flow economies of scale. *Transportation Research Part B, 32*(8), 605–616.

Racunica, I., & Wynter, L. (2005). Optimal location of intermodal freight hubs. *Transportation Research Part B: Methodological, 39*(5), 435–477.

Ramaswami, V., & Lucantoni, D. (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.

Stephan, F. (1958). Two queues under preemptive priority with Poisson arrival and service rates. *Operations Research, 6*(3), 399–418.

Uster, H., & Agrahari, H. (2011). A benders decomposition approach for a distribution network design problem with consolidation and capacity considerations. *Operations Research Letters, 39*(2), 138–143.