Yıldız, Barış; Savelsbergh, Martin W. P.; Dogru, Ali K.

Article

Transshipment network design for express air cargo operations in China

EURO Journal on Transportation and Logistics (EJTL)

Provided in Cooperation with:
Association of European Operational Research Societies (EURO), Fribourg

Suggested Citation: Yıldız, Barış; Savelsbergh, Martin W. P.; Dogru, Ali K. (2023) : Transshipment network design for express air cargo operations in China, EURO Journal on Transportation and Logistics (EJTL), ISSN 2192-4384, Elsevier, Amsterdam, Vol. 12, Iss. 1, pp. 1-14, https://doi.org/10.1016/j.ejtl.2023.100120

This Version is available at:
https://hdl.handle.net/10419/325190

EURO Journal on Transportation and Logistics 12 (2023) 100120

Contents lists available at ScienceDirect

EURO Journal on Transportation and Logistics

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

Transshipment network design for express air cargo operations in China

Barış Yıldız [a,*], Martin Savelsbergh [b], Ali K. Dogru [c]

[a] Department of Industrial Engineering, Koc University, Istanbul, Turkey
[b] H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[c] School of Management, University of Southern Mississippi, Hattiesburg, MS, USA

ARTICLE INFO

Keywords:
Transportation
Express air cargo
Transshipment network design
Multimodal transportation
Multiple service class
Integer programming

ABSTRACT

We introduce a novel multimodal (ground and air transportation) network design model with transshipments for the transport of express cargo with heterogeneous service classes (i.e., next morning delivery, and next day delivery). We formulate this problem using a novel path-based mixed-integer program which seeks to maximize the demand (weight) served. We investigate the value of the proposed transshipment network under various operational conditions and by benchmarking against a direct shipment network and a network with a single transshipment point which mimics a classical star-shaped hub-and-spoke network. Our extensive computational study with real-world data from ShunFeng (SF) Express reveals that the integration of ground and air transportation improves the coverage and that transshipment enables serving a large number of origin–destination pairs with a small number of cargo planes. Importantly, we show that by simplifying handling, i.e., employing cross-docking rather than time-consuming sortation, a transshipment network can transport express cargo fast enough to meet demanding delivery deadlines. Finally, we find that increasing the efficiency of intra-city operations and extending the nightly operating time window are the most effective operational adjustments for further improving the performance of the proposed transshipment network.

1. Introduction

The COVID-19 pandemic has revealed a need for resilient supply chains relying on responsive transportation networks that are less prone to delays and disruptions. Consequently, supply chains increasingly turn to express air cargo, as it provides safe, reliable, and fast delivery (and, at the same time, reduces the need to carry large inventories). As a result, the global air freight industry reported a record high revenue, reaching almost $128.8 billion in 2020 (IATA, 2021). This trend is expected to continue; global air cargo traffic is predicted to grow at 4% per year, and the global freighter fleet is projected to grow more than 60% until 2039 (Crabtree et al., 2020).

There are two common ways to transport express air cargo in practice: (1) using a star-shaped hub & spoke (HS) network, and (2) using a direct shipment (DS) network. We compare and contrast these common network types to a novel transshipment (TS) network that we propose in this paper; see Fig. 1 for an illustration. The first design, a star-shaped HS network, enables covering a large number of origin–destination pairs with a small number of cargo planes (no more than the number of origin and destination points). Despite its benefits, a star-shaped HS network has two major drawbacks. First, transporting air cargo via a single super hub requires sorting (i.e., unloading containers, sorting packages, and reloading containers), which is a time-consuming and costly (automated sortation equipment is expensive) operation and slows down the transportation of express air cargo as the time spent at the hub contributes significantly to the total transportation time. In certain settings, more complex hub network structures with multiple hubs can enhance efficiency (see, for example, Yıldız et al. (2021) and Wandelt et al. (2022)). However, for the express shipment application that we consider in this study, with next-morning and next-day service, such multi-hub network designs offer few advantages as most origin–destination paths involving more than one hub are not time-feasible because of the extra time required for sorting and inter-hub transfer. It is important to note that in a multi-hub design, the transfer of cargo between hubs can only begin once all the incoming cargo has been collected and processed (unpacked, sorted, and repacked). Similarly, final deliveries from the hubs can only commence once the inter-hub transfers have been completed and the packages have been processed again to be forwarded to their final destination (Yıldız et al., 2021). Besides the increased need for time-consuming sorting operations at hubs, such a system will be significantly slowed down by any long transfers in the access and hub networks, which is inevitable in a large geographic area such as China. Second, air cargo typically travels longer distances on a star-shaped HS network due to indirect trips. Increased travel distance not only delays deliveries

* Corresponding author.
E-mail address: byildiz@ku.edu.tr (B. Yıldız).

https://doi.org/10.1016/j.ejtl.2023.100120
Received 17 January 2023; Received in revised form 12 September 2023; Accepted 6 November 2023
Available online 10 November 2023

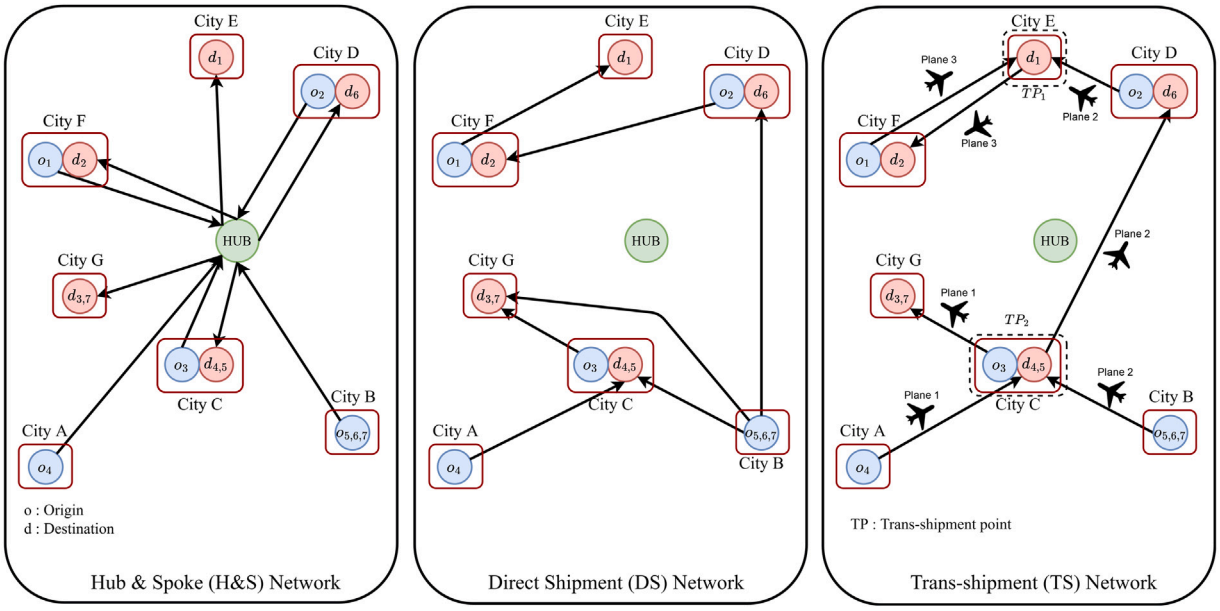

Fig. 1. Common air cargo transportation networks and the proposed transshipment network.

but also increases aircraft operating, maintenance, and repair costs. The second design, a DS network, allows for fast transportation of express air cargo since air cargo travels the shortest possible route on the network and does not require time-consuming intermediate sorting operations. However, a DS network requires acquiring and operating a large number of cargo planes (up to the number of origin–destination pairs) to cover a service region, which might not be feasible due to the high costs involved, particularly when there is a high number of origin–destination pairs to serve.

In this study, we propose a third design, a TS network in which cargo planes meet at transshipment points to exchange cargo. When we use the term transshipment, we refer to unloading an air cargo container from one aircraft and reloading that same air cargo container in another aircraft (a.k.a., cross-docking). Since packages with the same destination are loaded into the same containers, and only these containers (not the individual packages) are exchanged at the transshipment points without unloading, sorting, and reloading, handling is simpler in TS networks compared to sortation at a hub. Fig. 1 illustrates how a TS network operates. Note that to highlight the differences between the different networks, we have chosen not to use the hub as part of the DS and TS networks. However, in practice, the location of the hub may be such that it would make sense to include an airport at that location also in the DS and TS networks. In the TS network, we see that Plane 1 takes off from City A (Origin 4) with containers for City C (Destination 4). Plane 2 takes off from City B (Origins 5,6,7) with containers for City C (Destination 5), City D (Destination 6), and City G (Destination 7). Both planes go to the airport at City C, which is designated as a transshipment point (TP2). After containers for City C (Destinations 4,5) are unloaded from Plane 1 and 2 (these containers reach their destination), containers for City G (Destination 7) are transferred from Plane 2 to Plane 1. Plane 1 then flies to City G (Destination 7), and Plane 2 flies to City D (Destination 6). Similarly, at City E (transshipment point TP1), containers from Plane 2 are transferred to Plane 3, which then returns to City F (Destination 2) after delivering containers to City E (Destination 1). Here, we want to note that to keep this illustrative example simple, we did not consider the repeatability of the schedules (i.e., match the number of arriving and departing planes at a node), which we actually enforce in our model in Section 3.

Transshipment saves time compared to an HS network by eliminating the time-consuming complex sorting operations and enabling shorter travel distances. The advantage of a TS network over a DS network is that it can cover a set of origin–destination (OD) pairs with a smaller fleet. For applications where the number of OD pairs is too large for a DS network and too small to justify an HS structure, a TS network may be a suitable option. A TS network can also be used in conjunction with an HS network, thereby reducing the load of the sortation center and decreasing the average distances that packages travel.

Hybrid systems that use multiple network designs are not uncommon in practice. Many air cargo companies already use both a DS network and an HS network. In such a hybrid system, cargo is split into two groups: (1) large items that require time-definite service but that cannot be handled by automated sortation equipment because of their size, shape, or sensitivity (e.g., large machinery and equipment, and hazardous, chemical, or elevated temperature materials), and (2) small packages (e.g., online retail orders by customers) and medium-sized items (e.g., small machinery and equipment, small appliances, and medications) that can be carried in airfreight containers (unit loading device — ULD) and can be handled by automated sortation equipment. The first type of cargo is transported using a DS network with company-owned aircraft (i.e., belly capacity for these types of items on commercial flights is not an option because of space and handling requirements), whereas the second type of cargo is transported using an HS network, either using company-owned aircraft or relatively costly belly capacity on passenger aircraft. As mentioned earlier, the demand for the air transport of small- and medium-sized items is expected to continue to grow, overloading HS networks and necessitating expensive capacity investments (e.g., expanding sortation hubs and purchasing additional automated sortation equipment).

Introducing a TS network that complements an existing HS network can provide a less costly alternative. The demand for many origin–destination (OD) pairs is likely too small to fill a standard-size air cargo container, and replacing sortation with transshipment is not an option, but for some OD pairs, the demand can be large enough to fill one or more air cargo containers, which can then be trans-shipped. A TS network might be used to serve that demand, while the remaining demand is served using the HS network. This way, the introduction of a TS network takes some of the burden from the HS network and shortens the distance traveled for demands transshipped. Diverting a subset of demand to a TS network will not only speed up the handling and transportation of the items transshipped but may also speed up the handling and transportation of the items served by the HS network

because sortation time at the hub may be reduced, resulting in shorter wait times for aircraft. In addition to relieving the carrier from the burden of making significant investments in real estate and sortation equipment, the introduction of a TS network may offer the opportunity to enhance service offerings between some OD pairs, increase customer satisfaction, and reduce operating and transportation costs. Quantifying the value of introducing a TS network in a large existing air cargo transportation network is challenging and beyond the scope of our research. In this study, we assume that the subset of OD pairs to be served by the TS network has already been determined, and the goal is to design the TS network (i.e., designate the transshipment points and determine the cargo plane routes) and analyze its performance under various operational conditions (by benchmarking its performance against a DS network and a TS network with a single transshipment point, which resembles a star-shaped HS network).

Any effective express air cargo transportation network design has to consider two important features that complicate modeling and planning. The first one relates to the multimodal nature of express cargo transportation. In addition to the air cargo network itself, a ground transportation feeder network is operated as well. Coordinating these two networks effectively is critical to be able to deliver competitive service offerings in a large number of markets. The second one relates to the multiple service offerings, in terms of expediency of the delivery (e.g., next morning, next day by noon, next day by 6 p.m.). Capacity allocations in the two networks (ground and air) have to accommodate the individual requirements of the multiple service offerings. Consequently, designing a TS network that effectively serves a relatively small number of markets (OD pairs) that cover a relatively large geographical area and that can daily fill one or more airfreight containers requires a sophisticated and integrated approach.

To address the above-mentioned network design problem, we propose a novel transshipment-enabled, multimodal shipment model with differentiated service classes, and present a novel path-based formulation for it. Decomposing the cargo plane itineraries into one pickup and one delivery path, the proposed model allows this challenging network design problem to be formulated as a compact mixed integer program that can be solved directly for the real-world problem instances we consider in this study. Partnering with ShunFeng (SF) Express, the second-largest Chinese courier that offers domestic and international express delivery services, we conduct an extensive computational study based on real data to test the efficacy of our model and derive important managerial insight. Our experiments show that jointly planning ground and air transportation leads to increased demand coverage and that the use of transshipment enables serving a high number of OD pairs with a relatively small number of cargo planes. Moreover, we demonstrate that a network design that relies on a few transshipment points, which simplifies handling, eliminates time-consuming sortation, and reduces average travel distances (times), can cost-effectively support the competitive service offerings demanded by today's market. Finally, we find that increasing the efficiency of intra-city operations and extending the nightly operating time window are the most effective operational adjustments for further improving the performance of the proposed TS network.

The rest of the paper is organized as follows. Section 2 provides a review of recent and relevant literature. Section 3 defines the transshipment network design problem and introduces a novel mixed-integer programming formulation. Section 4 reports and analyzes the results of our computational study. Section 5 concludes the paper by summarizing our findings, discussing managerial insights, and making recommendations on possible future research avenues.

Before continuing, we want to emphasize that although the transshipment network design problem and the proposed solution approach are motivated by a collaboration with SF Express, the parameter settings and constraints used in our computational experiments do not provide an exact and complete representation of the operational environment at the company. While running certain what-if scenarios, we deliberately alter some parameters and constraints to capture and highlight distinctive characteristics of the problem and derive critical managerial insight.

2. Literature review

Express delivery network design is a well-established literature that dates back to the 1990s. Providing a comprehensive review is beyond our scope, so we refer the reader to in-depth reviews by Crainic (2000), Agarwal (2002), Prodhon and Prins (2014), SteadieSeifi et al. (2014), Feng et al. (2015), and more recently by Alumur et al. (2021). Table 1 compares and contrasts relevant modeling studies according to their focus (network design, pickup and delivery, and air cargo scheduling), transportation mode (single vs. multiple), network type (hub & spoke with a single hub, hub & spoke with multiple hubs, direct shipment, and multiple transshipment points), other factors considered (time constraint, and service levels), and solution methodology.

Studies that focus on network design (ND) seek to locate a single or multiple hubs, while simultaneously determining the allocation of non-hub nodes to hub-nodes to serve origin–destination pairs so that either a network-wide profit is maximized or a total cost function is minimized. As can be seen, the majority of ND studies rely on hub & spoke networks. While earlier works, such as Kim et al. (1999), Armacost et al. (2002), Barnhart et al. (2002), Smilowitz and Daganzo (2007), consider a single hub, more recent studies, such as Lin and Chen (2008), Yaman (2009), Alumur et al. (2012b,a), Louwerse et al. (2014), Serper and Alumur (2016), Yu et al. (2017), Lee et al. (2019), Yıldız et al. (2021), Dai et al. (2021), and Wandelt et al. (2022) consider multiple hubs. These studies typically do not make a distinction among delay-tolerant parcels and time-sensitive parcels with special packaging, labeling, handling, and spacing requirements. While hub & spoke networks with multiple hubs might be more suitable for delay-tolerant parcels that can be transported in the belly of passenger aircraft, star-shaped networks (hub & spoke networks with a single hub) and direct shipment networks (DS) are more effective strategies for transporting time-sensitive parcels and critical items (Büdenbender et al., 2000; Lin and Chen, 2003; Cohn and Barnhart, 2003; Yan et al., 2006; Yan and Chen, 2008; Derigs et al., 2009; Qu and Bard, 2012; Derigs and Friederichs, 2013; Yıldız and Savelsbergh, 2022). Our study essentially extends the former approach, by establishing multiple star-shaped networks, such that each hub in the center of a star-shaped network functions as a transshipment point. Unlike hub networks, the transshipment points in our model are not connected and avoid loss of time due to inter-hub transfers which require complex and time-wise costly coordination of cargo that belongs to different origin–destination pairs.

Several modeling studies focus on air cargo scheduling (ACS) to find cyclic timetable schedules for cargo aircraft (Yan et al., 2006; Yan and Chen, 2008; Derigs et al., 2009; Derigs and Friederichs, 2013; Louwerse et al., 2014). Extending the basic scheduling problem for the cargo, Qu and Bard (2012) solves a pickup and delivery (PUD) problem with a fixed number of identical vehicles with limited capacities that need to be routed among depots while respecting hard time window constraints. Similar to ours, these studies consider multi-leg transfers for the cargo. However, while we focus on the design of the network, these studies focus on the flow of commodities along existing links (air and ground) provided by preplanned cargo plane flights and purchased belly capacity on commercial passenger flights.

A major complicating factor is the time-constrained nature of express air cargo operations, which is considered by most papers examined. These studies either use time-extended networks or location–allocation models with time windows to capture flight time interdependencies. The multimodal nature of express shipment operations, on the other hand, is somewhat under-explored. Few studies consider air and ground transportation (Kim et al., 1999; Barnhart et al., 2002; Smilowitz and Daganzo, 2007; Alumur et al., 2012b; Serper and Alumur, 2016; Yıldız and Savelsbergh, 2022), whereas Dai et al. (2021) consider air and rail transportation. Considering air and ground transportation simultaneously typically increases computational complexity, so understandably, modelers often isolate an air cargo network from

Table 1
Classification of relevant literature.

Study	Focus	Modes	Network type				Service		Methodology
			HS (single)	HS (multiple)	DS	MTP	Time Constr.	Service level	
Kim et al. (1999)	ND	Multiple	✓				✓	Multiple	MIP, CG, RG
Budenbender et al. (2000)	ND	Single			✓		✓	Single	TS, BB
Armacost et al. (2002)	ND	Single	✓				✓	Single	LP
Barnhart et al. (2002)	ND	Multiple	✓				✓	Multiple	MIP, BPC , BP
Lin and Chen (2003)	ND	Single				✓		Single	MIP, BB
Yan et al. (2006)	ACS	Single				✓	✓	Single	MIP
Smilowitz & Daganzo (2007)	ND	Multiple	✓				✓	Multiple	CA
Cohn et al. (2008)	ND	Multiple			✓			Single	IP
Lin & Chen (2008)	ND	Multiple		✓	✓		✓	Single	DFS
Yan and Chen (2008)	ACS	Single				✓	✓	Single	MIP
Derigs et al. (2009)	ACS	Single				✓	✓	Single	CG, SP
Yaman (2009)	ND	Single		✓				Single	MIP
Alamur et al. (2012a)	ND	Multiple		✓			✓	Single	MIP
Alamur et al. (2012b)	ND	Multiple		✓			✓	Multiple	MIP, Heur.
Qu & Bard (2012)	PUD	Single				✓	✓	Single	GRASP
Derigs & Friederichs (2013)	ACS	Single				✓	✓	Multiple	MIP, BPC
Louwerse et al. (2014)	ACS	Multiple		✓			✓	Multiple	CG, LS
Serper & Alamur (2016)	ND	Multiple		✓				Single	MIP, VNS
Yu et al. (2017)	ND	Single		✓				Single	MIP, GA, FW
Lee at al. (2019)	ND	Single		✓				Single	SIB
Yildiz et al. (2021)	ND	Multiple		✓			✓	Single	MIP
Yildiz & Savelsbergh (2021)	ND	Multiple			✓		✓	Multiple	MIP, CG
Dai et al. (2022)	ND	Multiple		✓				Single	CPn4, MMHUBBI
Wandelt et al. (2022)	ND	Single	✓	✓				Single	BD, RG, CPn3, CPn4
Our study	ND	Multiple				✓	✓	Multiple	MIP

MTP: Multiple transshipment points, ND: Network Design, PUD: Pick Up & Delivery, ACS: Air Cargo Scheduling, HS: Hub & Spoke, DS: Direct Shipment, RG: Row Generation
CA: Continuous Approximation, IP: Integer Programming, LP: Linear Programming, MIP: Mixed Integer Programming, BB: Branch & Bound, BD: Benders Decomposition
BP: Branch & Price, BPC: Branch & Price & Cut, CG: Column Generation, DFS: Depth First Search, LS: Local Search, SP: Shortest Path, VNS: Variable Neighborhood Heuristic
SIB: Swarm Intelligence Based Heuristic, TS: Tabu Search, GA: Genetic Algorithm, FW: Frank-Wolfe Algorithm, MMHUBBI:Multi-modal Iterative ND Algorithm
CPn3:CPLEX Cubic, CPn4:CPLEX Quadratic

a ground transportation network. However, as we will show there is a significant benefit in modeling these two complementary transportation modes together as it will extend service coverage in a cost-efficient way. The transshipment option, which is another important factor that allows for greater flexibility and cost reduction due to better utilization of resources, has also attracted relatively little attention in the literature. We could identify only 6 papers that allow transshipment: namely Lin and Chen (2003), Yan et al. (2006), Yan and Chen (2008), Derigs et al. (2009), Qu and Bard (2012) and Derigs and Friederichs (2013). As mentioned before most of these studies do not have a network design focus. Finally, express air cargo companies typically offer multiple service classes, each with a different delivery time promise, such as next-morning, next-day, or 2-day delivery. However, considering multiple service classes adds another layer of complexity to an already difficult problem. Studies that examine the impact of multiple service classes can be found in Kim et al. (1999), Barnhart et al. (2002), Smilowitz and Daganzo (2007), Alumur et al. (2012a), Derigs and Friederichs (2013), Louwerse et al. (2014), Yıldız and Savelsbergh (2022).

Yu et al. (2017), and Yıldız and Savelsbergh (2022) closely align with our study, as both studies also investigate the SF Express network in China. However, many differences still exist. Yu et al. (2017) determine multiple hub locations and allocations between non-hub and hub nodes, using company-owned air cargo capacity and renting additional belly capacity from passenger airlines, via a bi-level model. They use a genetic algorithm to solve the upper model, which selects hubs and generates links, and a Frank–Wolfe algorithm to solve the lower model, which simulates solution qualities of chosen paths to determine the best solution. However, they assume a single service class, and they neither consider ground transportation nor use time constraints. Yıldız and Savelsbergh (2022) use column generation to solve a direct shipment model that considers company-owned air cargo and ground transportation to distribute time-sensitive parcels with multiple service classes. However, their model does not consider transshipment. Our study differs from these two studies and others in that we solve a complex transshipment network design problem with multiple transportation modes (ground and company-owned air cargo fleets) used to deliver time-sensitive items with multiple service classes under spatial and temporal constraints to maximize demand coverage. Since demand for these time-critical items is relatively sparsely and evenly distributed among a small number of demand points, we present a novel path-based MIP formulation that relies on concatenations of feasible segments to generate and evaluate the solution quality of paths. We show the value of our model via extensive computational experiments based on real data. In the next chapter, we formally define this problem, introduce key notation, present our model formulation, and discuss unique operational characteristics.

3. Problem definition

3.1. Operational environment

Consider an express package carrier that operates an HS network with a single (central) sortation hub. To complement this HS network, the carrier seeks to introduce a transshipment network so as to ease the burden on the sortation hub and to decrease the average air-transportation distance per unit load. As shipments that can possibly

Table 2
Key notation.

Sets		Decision variables	
Q	Set of demands	t_h	The latest arrival time of pick-up path at transshipment
C	Set of cities		location $h \in H^T$
A	Set of cities with an airport	u_p	Binary, 1 if pick path $p \in \mathcal{P}$ is used, 0 otherwise
H	Set of hub cities	$u_{\bar{p}}$	Binary, 1 if delivery path $\bar{p} \in \bar{\mathcal{P}}$ is used, 0 otherwise
H^T	Set of alternative transshipment locations	x_p^q	Binary, 1 if demand $q \in Q$ is served on path $p \in \mathcal{P} \cup \bar{\mathcal{P}}$,
P	Set of all feasible actual pick-up paths		0 otherwise
$\bar{P}$	Set of all feasible actual delivery paths	y^q	Binary, 1 if demand $q \in Q$ is served, 0 otherwise
P^o	Set of empty pick up paths	z_h^q	Binary, 1 if demand $q \in Q$ is transferred at transshipment
$\bar{P}^o$	Set of empty delivery paths		location $h \in H^T$, 0 otherwise
$\mathcal{P}$	Set of all feasible pick-up paths	e_i^q	Binary, 1 if demand $q \in Q$ is served with an empty pick up
$\bar{\mathcal{P}}$	Set of all feasible delivery paths		path at transshipment location $i \in H^T$, 0 otherwise
$\mathcal{P}(q)$	Set of all feasible pickup paths for a demand $q \in Q$	e_i^{-q}	Binary, 1 if demand $q \in Q$ is served with an empty delivery
$\bar{\mathcal{P}}(q)$	Set of all feasible delivery paths for a demand $q \in Q$		path at transshipment location $i \in H^T$, 0 otherwise
Q_p	Demands that can be assigned to a pickup path $p \in \mathcal{P}$		
$\bar{Q}_p$	Demands that can be assigned to a delivery path $p \in \bar{\mathcal{P}}$		
Parameters			
o^q	Origin of demand $q \in Q$	t_{ij}^g	Ground travel time from city i to city j, $i, j \in C$
d^q	Destination of demand $q \in Q$	τ	Time required for freight handling operations
f^q	Weight of demand $q \in Q$		at transshipment points
t^q	Ready time of demand $q \in Q$	e_p^q	Lower bound on the arrival time of a plane
$\bar{t}^q$	Due time of demand $q \in Q$		executing pick up path p at the transshipment
κ	Number of demands (pallets) a plane can carry. is served		location d_p, if it serves demand q
η	Number of company-owned cargo planes	e_p^{-q}	Time required for demand q to reach its
$\underline{\phi}$	Start time of period during which cargo planes		destination d_q, if it is served on delivery path $\bar{p} \in \bar{P}$
	can operate	δ_p	Duration of path $p \in P$
$\overline{\phi}$	End time of period during which cargo planes	a_i	Segment i on path $p \in P$
	can operate	σ	Percentage of shipments to be delivered next morning before noon
t_{ij}^a	Air travel time from city i to city j, $i, j \in A$		

be served by the transshipment network, the carrier considers pallets (air-cargo containers) composed of inter-city shipments with the same origin, destination, ready time, and service class. We refer to each such pallet as a demand in our model. The weight of a demand is the sum of the weights of the shipments that constitute the demand. The due time at the destination represents the time by which the shipments need to reach the inbound gateway hub of the destination city to be able to be delivered in time to meet service class requirements. In defining due dates, as is done in Yıldız and Savelsbergh (2022), we assume that the time it takes to perform intra-city operations is known, i.e., the time between when a package is picked up and the time the package becomes available for transport at a city's outbound gateway hub as well as the time between the arrival of a package at a city's inbound gateway hub and the time it is dropped off at its final destination are known and the same for every package.

The express package carrier reserves a number of its aircraft for the transshipment network. Due to government regulations, these cargo planes can only be operated during a given time window; between 11 pm and 8 am in China. Moreover, updates to aircraft schedules are subject to coordination with the aviation authorities and thus cannot be dynamically changed according to the demand realizations. We refer to cities where the carrier has airport operations, i.e., freight handling, as *hub cities*, and all other cities as *non-hub cities*. If beneficial, a combination of ground and air transportation can be used when serving a demand, for example, when the origin and/or destination city (or both) are non-hub cities.

Given that serving demand using a transshipment network can be beneficial (i.e., require less per-unit air transfer distance and save on sorting time and capacity), the express package carrier seeks to design a transshipment network that can support the maximum demand, in terms of total weight, given the number of aircraft the company dedicates to the transshipment network. The design of the network involves determining the transshipment airports, the demands to serve, and the routes and schedules of the aircraft. Observe that we are focusing on the design of the air service network operations, i.e., the aircraft routes and schedules. The demand can be based on historical data but may be adjusted to reflect market trends or to account for anticipated growth as a result of pricing policy changes, etc. We want to emphasize that in this study we assume that service coverage and pricing decisions, and thus the total revenue, are given (i.e., these strategic decisions have already been made), hence profit maximization is equivalent to cost minimization, which is achieved through the optimal design of the transshipment network.

3.2. Notation

Let Q be the set of demands. Each demand $q \in Q$ is characterized by a 5-tuple $\langle o^q, d^q, f^q, t^q, \bar{t}^q \rangle$, where o^q is the demand's origin, d^q its destination, f^q its weight, t^q its ready time, and $\bar{t}^q$ the due time. Let the set of cities, the set of cities with an airport, the set of hub cities, and the set of hub cities where transshipment can take place be denoted by C, A, H, and H^T, respectively. Between any two cities $i, j \in C$, let t_{ij}^g denote the ground travel time from i to j. Similarly, between any two cities with an airport $i, j \in A$, let the air travel time from i to j be denoted by t_{ij}^a. The travel times are assumed to include any time required for handling operations at the gateway hubs. The number of cargo planes is denoted by η. The capacity of a plane, i.e., the number of demands (pallets) it can carry, is denoted by κ. The cargo planes are only allowed to operate between $[\underline{\phi}, \overline{\phi}]$. Next, we present details of the transshipment model and our solution approach. For a quick reference, Table 2 presents the notation used throughout the paper.

3.3. Transshipment model (TSM)

In the transshipment model (TSM), a demand transported by air may occupy more than one flight leg, often, but not necessarily, on more than one cargo plane. That is, cargo planes are allowed to meet in a transshipment location to exchange some or all of their cargo. Conceptually, each demand uses one pickup and one delivery flight that meet at a transshipment location. If the transshipment location happens to be the hub-city where the demand enters or exits the air network, then the demand uses only a single flight (either a pickup or a delivery flight). In the TSM, each cargo plane's route is composed of two distinct parts: a pickup flight and a delivery flight (where one of them may be the "empty" flight). On the pickup flight, the cargo plane collects demands at various hub-cities to take them to a transshipment location. On the delivery flight, the cargo plane takes demands from the transshipment location and drops them off at various hub-cities. Ground transportation can be used to transfer demand from/to a non-hub city to/from a hub-city. The TSM seeks to make the most effective use of the flying time available for cargo planes to provide broad coverage, i.e., it seeks to provide connections between a large number of cities (possibly at the expense of reduced capacity utilization). It is important to note that the TSM only considers *transfers*, i.e., cargo can be transferred from one plane to another at certain designated locations, but no sorting takes place at these locations.

We assume that for the fixed number of cargo planes allocated to the transshipment network, the express carrier seeks to transport as much demand as possible (in terms of total weight) in the transshipment network to reduce the average distance demand travels and the need for costly sorting operations.

We further assume that both the pick-up and the delivery flights are simple paths, i.e., a hub-city is visited at most once on a pick-up or a delivery path (we will show that this is not a restrictive assumption and that there always exists an optimal solution that only uses simple paths). A pick-up path can be represented by a sequence of flight legs, $p = (a_1, \dots, a_m)$ with $a_i \in H$, for $i < m$, and $a_m \in H^T$. Similarly, a delivery path can be represented by a sequence of flight legs $p = (a_1, \dots, a_m)$ with $a_1 \in H^T$ and $a_i \in H$, for $i > 1$.

The duration of a path p is denoted by δ_p. We only consider pick-up and delivery paths that start and end in the allowed operating window $[\underline{\phi}, \overline{\phi}]$ and we refer to such paths as *feasible*. A pickup or delivery path $p = (a_1, a_2, \dots, a_m)$ is called an *actual-path* if $m \geq 1$ and we denote the set of all feasible actual pick-up and delivery paths with P and $\bar{P}$, respectively. For technical reasons that will become clear soon, we also consider *empty paths*, i.e., we allow $m = 0$. The set of empty pick-up and delivery paths are denoted by P^o and $\bar{P}^o$, respectively. We define $\mathcal{P} = P \cup P^o$ as the set of all feasible pick-up paths and $\bar{\mathcal{P}} = \bar{P} \cup \bar{P}^o$ as the set of all feasible delivery paths.

For each demand $q \in Q$, let e_p^q denote a lower bound on the arrival time at transshipment location d_p of a plane executing pickup path p ending at transshipment location d_p if it serves demand q. Note that the calculation of e_p^q involves determining the hub-city from which the demand q is going to be picked up (where the demand enters the air transportation network) and the time required by ground transportation from o^q to that hub-city. Furthermore, let $\bar{e}_{\bar{p}}^q$ denote the time it takes for demand q to reach its destination d^q (including any ground transportation), if it is served on delivery path $\bar{p} \in \bar{P}$. Note that the calculation of $\bar{e}_{\bar{p}}^q$ involves determining the hub-city where demand q will leave the air transportation network to reach its final destination.

Let τ denote the time required to transship pallets from one plane to another at a transshipment location. For each demand $q \in Q$, a pickup path p is called a *plausible* pickup path if there exists a delivery path $\bar{p} \in \bar{\mathcal{P}}$ such that $d_p = o_{\bar{p}}$ and $e_p^q + \tau + \bar{e}_{\bar{p}}^q \leq t^q$. A *plausible* delivery path for a demand $q \in Q$ is defined similarly. The sets of all plausible pickup and delivery paths for demand $q \in Q$ are denoted by $\mathcal{P}(q)$ and $\bar{\mathcal{P}}(q)$, respectively. Similarly, we define Q_p and $\bar{Q}_{\bar{p}}$ as the set of demands that can be assigned to a pickup path $p \in \mathcal{P}$ and a delivery path $\bar{p} \in \bar{\mathcal{P}}$, respectively.

In the TSM, the movement of a package from its origin to its destination is thought of as consisting of four segments. In the first segment, ground transportation is used to reach the location where the package enters the air transportation network. In the second segment, a pickup path is used to take the package to a transshipment location. In the third segment, a delivery path is used to take the package to a location where it exits the air transportation network. In the final segment, ground transportation is used to reach the package's final destination. The TSM seeks to find feasible concatenations of segments for each demand so as to maximize the demand, in terms of total weight, that can be supported by the transshipment network. Concatenation of segments, however, requires both spatial and temporal coordination to ensure that feasible paths are constructed for the demands. The following observations allow us to construct feasible concatenations without explicitly modeling the ground transportation segments by incorporating their timing information in the pickup and delivery paths.

Observation 1. *Let $p \in \mathcal{P}$ be a pick-up path and let $\hat{Q} \subset Q$ be the set of demands served by p, then the earliest time t^* that a plane executing p can arrive at d_p is $\max_{q \in \hat{Q}} \{e_q^p\}$.*

Observation 2. *Let $\bar{p} \in \bar{\mathcal{P}}$ be a delivery path, let $\hat{Q} \subset Q$ be the set of demands served by $\bar{p}$, and let the earliest departure time of a plane executing delivery path p be t^* (implied by the latest arrival time of a demand in $\hat{Q}$ and the transshipment time τ), then each demand $q \in \hat{Q}$ can arrive at its destination at time $t^* + \bar{e}_{\bar{p}}^q$.*

Observation 3. *Let p be a non-simple pick-up path and let $\hat{Q} \subset Q$ be the set of demands served by p. Then there exists a simple path $\hat{p} \in \mathcal{P}$ that visits the same set of locations, picks up the same set of demands, and arrives at the transshipment location at the same time as path p.*

This follows from the fact that whenever a location is visited more than once, it is feasible to pick up all demand at that location at the time of the last visit.

Observation 4. *Let p be a non-simple delivery path and let $\hat{Q} \subset Q$ be the set of demands served by p. Then there exists a simple path $\hat{p} \in \mathcal{P}$ that visits the same set of locations, delivers the same set of demands, and departs the transshipment location at the same time as path p.*

This follows from the fact that whenever a location is visited more than once, it is feasible to deliver all demand at that location at the time of the first visit.

Next, we present a mixed-integer programming formulation for the TSM. We introduce the following decision variables:

- *Time-coordination* variables $t_h, h \in H^T$, indicating the latest arrival of pickup paths at transshipment location $h \in H^T$;
- *Pickup-path* variables $u_p, p \in \mathcal{P}$, indicating whether a pickup path is used or not;
- *Delivery-path* variables $u_{\bar{p}}, \bar{p} \in \bar{\mathcal{P}}$, indicating whether a delivery path is used or not;
- *Demand assignment* variables x_p^q, $p \in \mathcal{P} \cup \bar{\mathcal{P}}$, indicating whether a demand $q \in Q$ is served on path $p \in \mathcal{P} \cup \bar{\mathcal{P}}$ or not;
- *Coverage* variables y^q, $q \in Q$, indicating whether a demand q is served or not;
- *Transshipment* variables z_h^q, $q \in Q, h \in H^T$, indicating whether a demand q is transferred at transshipment location $h \in H^T$ or not.
- *Empty-path* variables ϵ_i^q / $\bar{\epsilon}_i^q$, $i \in H^T$, indicating whether a demand $q \in Q$ is served with an empty pickup/delivery path at transshipment location i or not.

With these decision variables, formulation TS is defined as follows:

$$\max \sum_{q\in Q} f^q y^q \tag{1}$$

$$\text{s.t.} \sum_{h\in H^T} z_h^q \geq y^q \qquad \forall q \in Q \tag{2}$$

$$\sum_{\substack{p\in P(q)\\ d_p=h}} x_p^q + \epsilon_h^q = z_h^q \qquad \forall q \in Q, \forall h \in H^T \tag{3}$$

$$\sum_{\substack{\bar{p}\in \bar{P}(q)\\ o_{\bar{p}}=h}} x_{\bar{p}}^q + \bar{\epsilon}_h^q = z_h^q \qquad \forall q \in Q, \forall h \in H^T \tag{4}$$

$$t_h \geq \sum_{\substack{p\in P(q)\\ d_p=h}} e_p^q x_p^q + t^g_{o^q,h}\epsilon_h^q \qquad \forall q \in Q, \forall h \in H^T \tag{5}$$

$$t_h + \tau + \sum_{\substack{\bar{p}\in \bar{P}(q)\\ s_{\bar{p}}=k}} \bar{e}_{\bar{p}}^q x_{\bar{p}}^q + t^g_{h,d^q}\bar{\epsilon}_h^q \leq \bar{l}^q \qquad \forall h \in H^T, \forall q \in Q(h) \tag{6}$$

$$t_{o_{\bar{p}}} + \tau + \delta_{\bar{p}} u_{\bar{p}} \leq \overline{\phi} \qquad \forall \bar{p} \in \bar{P} \tag{7}$$

$$\sum_{p\in\mathcal{P}} u_p = \rho \tag{8}$$

$$\sum_{q\in Q_p} x_p^q \leq \kappa u_p \qquad \forall p \in \mathcal{P} \tag{9}$$

$$\sum_{q\in Q_p} x_{\bar{p}}^q \leq \kappa u_{\bar{p}} \qquad \forall \bar{p} \in \bar{\mathcal{P}} \tag{10}$$

$$\sum_{\substack{p\in\mathcal{P}\\ d_p=h}} u_p - \sum_{\substack{\bar{p}\in\bar{\mathcal{P}}\\ s_{\bar{p}}=h}} u_{\bar{p}} = 0 \qquad \forall h \in H^T \tag{11}$$

$$\sum_{\substack{p\in\mathcal{P}\\ s_p=h}} u_p - \sum_{\substack{\bar{p}\in\bar{\mathcal{P}}\\ d_{\bar{p}}=h}} u_{\bar{p}} = 0 \qquad \forall h \in H^T \tag{12}$$

$$t_h \geq 0 \qquad \forall h \in H^T \tag{13}$$

$$u_p \in \{0,1\} \qquad \forall p \in \mathcal{P}\cup\bar{\mathcal{P}} \tag{14}$$

$$x_p^q \in \{0,1\} \qquad \forall q \in Q, \forall p \in \mathcal{P}\cup\bar{\mathcal{P}} \tag{15}$$

$$y^q \in \{0,1\} \qquad \forall q \in Q \tag{16}$$

$$z_h^q, \epsilon_h^q, \bar{\epsilon}_h^q \in \{0,1\} \qquad \forall q \in Q, h \in H^T \tag{17}$$

The objective is to maximize the demand served (in terms of total weight). Constraints (2)–(4) ensure that if a demand $q \in Q$ is served, it is served by a pick-up and a delivery path (including empty paths) that arrive at and depart from the same transshipment location. Note that Constraints (2)–(4) ensure spatial coordination, while Constraints (5)–(6) ensure that the timing of the paths is synchronized and demand q is delivered on time. Constraints (7) ensure that cargo planes respect the operating window. Constraint (8) limits the number of planes used in the air transportation network, and Constraints (9) and (10) ensure that the number of demands assigned to an operated path is not more than the available capacity. Finally, Constraints (11) and (12) enforce that the number of arriving and departing planes in a transshipment location match, and the resulting schedule can be repeated on a daily basis.

Note that considering the pick-up and delivery segments executed by a plane separately, as opposed to considering complete itineraries for a plane, significantly reduces the number of variables. The observations above show that time coordination of pick-up and delivery segments can be achieved without introducing timed copies of path segments (with different visit times at the locations), which also significantly reduces the number of variables. In fact, preliminary experiments have shown, mostly due to the limited cargo plane operating window (nine hours in practice), that it is possible to solve TS directly without the need for more involved approaches, such as column generation or branch and price. To improve the computational efficiency further, we added the following valid inequalities to the model, which improved the computational efficiency in our preliminary experiments.

$$x_p^q \leq u_p \qquad \forall q \in Q, \forall p \in P_q \tag{18}$$

$$x_{\bar{p}}^q \leq u_{\bar{p}} \qquad \forall q \in Q, \forall \bar{p} \in \bar{P}(q) \tag{19}$$

Validity of the inequalities (18) and (19) simply follows from the observation that both the assignment and the path variables are binary decision variables and a demand can be assigned to a pickup or delivery path only if the path is executed. For the problem instances we study, the total number of inequalities in (18) and (19) are not prohibitively large and the additional computational effort needed to solve the linear relaxation with a higher number of constraints is offset by the increased efficiency in the branch and bound search. Similarly, we also tested tightening the capacity constraints (9) and (10) by replacing the plane capacity κ with a path-specific capacity $\kappa_p = min\{\kappa, |Q_p|\}$ for a pickup or delivery path p. However, our preliminary experiments have shown that for almost all paths in our problem instances, we have $\kappa_p = \kappa$ and, thus, there is no practical benefit of such a tightening.

4. Computational study

In this section, we present the results of a comprehensive numerical study based on SF Express data (data the company uses in its strategic network analysis efforts). First, we provide an overview of the data to give the reader a flavor of the operational environment. Then, we outline our parameter setting, and we explain modifications of the original data to enable investigation of certain what-if scenarios. Our computational experiments seek to answer the following research questions:

- What characteristics does a high-quality transshipment air network design have (i.e., number and locations of airports where transshipments take place, pickup routes, delivery routes, ground transportation, etc.)?
- Does the model effectively integrate ground and air transportation? If yes, what value does multimodal transportation provide?
- What value, if any, does using transshipments provide to the network? What are the main challenges and limitations?
- What operational efficiency improvements have the greatest potential to improve system performance (i.e., improving the efficiency of intra-city and/or airport handling operations, modifying the cargo plane operation window, increasing the fleet size, etc.)?
- What is the impact of differentiated service class distributions on the optimal solutions?

To answer these research questions, we conduct experiments to establish (1) the value of multimodal transportation (i.e., ground and air) and of using transshipments using different fleet sizes, (2) the impact of changes in the operating environment (i.e., intra-city operations time, airport operations time, and operations time window), and (3) the impact of service class distribution. We believe that solving large problem instances in a reasonable amount of time is sufficient evidence of the efficacy of the proposed solution approach. Therefore, rather than conducting separate experiments to demonstrate the computational efficiency of our solution approach, we simply report the solution times and the optimality gaps as part of the computational results for our experiments. Finally, we discuss the managerial implications of the results, which may be of use to other express air cargo companies.

4.1. Data

In our experiments, we use the demand data that SF Express uses to make informed strategic decisions about its current network. However, to be able to answer the research questions we aim to answer in this study, we only consider demands (shipments consolidated by origin, destination, ready time, and service level) with a weight between 1000 and 2000 kg. Smaller demands are not included to ensure a minimum pallet fill rate (80%), and larger demands are not considered as they are better suited for direct shipment (Yıldız and Savelsbergh, 2022). Furthermore, we consider only those demands that require air

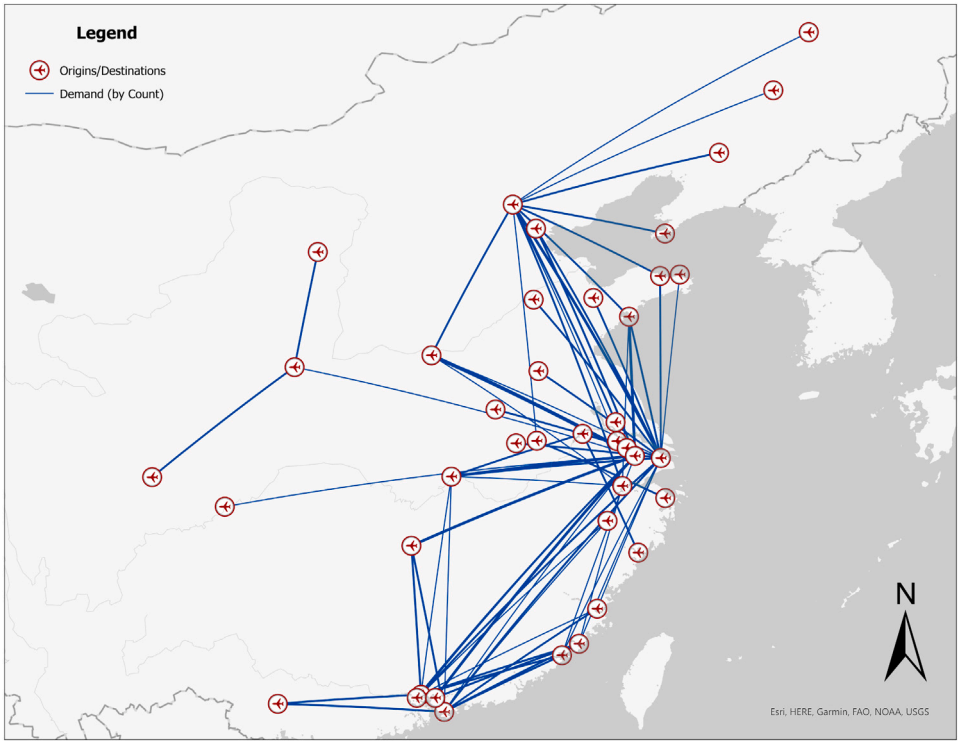

Fig. 2. The spatial distribution of demand.

transportation (i.e., if the ground transportation time is short enough to meet the delivery deadline, we assume ground transportation will be used). Finally, we only consider the demand for next-morning and next-day service offerings. Below we present details regarding the data used in our experiments.

- In the data, there are 65 demands that have weights between 1000 and 2000 kg and that require air transportation to meet next-day delivery deadlines. Based on discussions with SF Express, we assume 5.5 h for intra-city operations, for both pick-up (from pick-up locations to gateway hubs) and delivery (from gateway hubs to delivery locations). Considering these intra-city operation times, 56 of the demands are available for air-cargo operations before the cutoff time (midnight). Thirty-four of these demands are between hub-cities, and 22 require ground transportation (i.e., either the origin or the destination is not a hub-city). In Fig. 2, we illustrate the spatial distribution of the demands over the service region. Blue lines represent demands (origin–destination pairs), where the thicknesses of the lines are proportional to their weights.
- A demand has one of two service classes: (1) “next-day service” (i.e., delivery by 6 pm the next day), and (2) the “next-morning service” (i.e., delivery before noon the next-day). In Fig. 3, we show the cumulative ready time distribution of demands.
- We consider cargo planes with a 14-ton capacity. Chinese aviation regulations allow cargo planes to operate only from 11 p.m. to 8 a.m., which implies a 9-hour operating period.
- Ground transportation is available for origin–destination pairs with a travel time of less than six hours (based on travel times determined by SF Express). Note, however, that if no ground transportation option is available between two locations, service between these two locations may still be offered in the form of ground–air–ground as long as the origin and destination can both be reached in less than six hours by ground transportation from an air hub. We assume a flexible ground transportation system, which can be “aligned” with the express air network. Therefore, we assume that ground transportation between a non-airport and an airport city arrives at the gateway hub of an airport city at the arrival cut-off time (which accounts for the time required to load departing cargo planes). Similarly, we assume that ground transportation between an airport and a non-airport city departs from the gateway hub of an airport city at the available time (which accounts for the time required to unload arriving cargo planes).
- We assume that transferring demands between cargo planes at a transshipment location takes 1.25 h and that loading/unloading demands at pick-up/delivery stops takes 1 h.

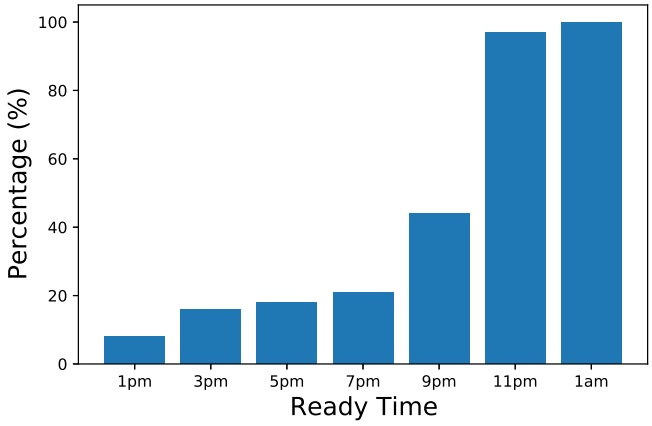

Fig. 3. Distribution of demand ready times.

Our preliminary computational experiments revealed that considering only the top eight cities in terms of in- and outbound demand as candidate locations for transshipment can significantly reduce the run times (approximately 10 times) with minimal loss in solution quality (less than one percent). Therefore, we limit the candidate transshipment locations accordingly.

Table 3
The impact of ground integration, fleet size, and transshipments.

Model	ground?	# planes	Count	Weight	#T	Time	Gap
TSM	yes	2	30	36,102	2	104	0.00%
TSM	no	2	14	17,075	2	3	0.00%
STP	yes	2	27	32,630	1	340	0.00%
DS	yes	2	39	46,226	–	6	0.00%
TSM	yes	3	40	47,825	3	12,810	0.00%
TSM	no	3	19	22,935	2	6	0.00%
STP	yes	3	37	44,693	1	4097	0.00%
DS	yes	3	41	49,109	–	6	0.00%
TSM	yes	4	48	56,845	3	43,202	1.60%
TSM	no	4	23	27,812	2	10	0.00%
STP	yes	4	46	54,717	1	1239	0.00%
DS	yes	4	45	53,718	–	3	0.00%
TSM	yes	5	51	60,129	3	585	0.00%
TSM	no	5	26	31,329	3	7	0.00%
STP	yes	5	49	57,752	1	2034	0.00%
DS	yes	5	48	57,064	–	2	0.00%
TSM	yes	6	51	60,129	2	320	0.00%
TSM	no	6	28	33,584	3	20	0.00%
STP	yes	6	50	59,100	1	57	0.00%
DS	yes	6	43	62,399	–	2	0.00%

The design problem we focus on is typically solved every 3 to 6 months to accommodate seasonal changes in demand. Due to Chinese aviation regulations, the plane routes (including the cargo planes) can only be changed with government approval, which makes it practically impossible to have frequent changes in the network design. As such, the run time of the solution approach is not a big concern. However, to complete our experiments in a reasonable amount of time, we impose a 12 h runtime limit, observing that beyond that time limit, no new feasible solutions are typically found, and the lower bound hardly changes.

We elaborate on other experiment-specific details in the following subsections.

4.2. Value of multimodal transportation and air cargo transport with transshipments

We measure the value of multimodal transportation by comparing the performance of a network that only uses air cargo capacity with a network that uses ground and air cargo capacities simultaneously. To analyze the benefits of transshipments, we consider two benchmarks:

- The Direct Shipment (DS) model proposed by Yıldız and Savelsbergh (2022) to transfer demands that are relatively densely distributed over the service region.
- The Single Transshipment Point (STP) model in which only one of the candidate airports is chosen as the transshipment point.

We also vary fleet size between two and six planes to understand the impact of fleet size on the comparative advantages of the considered approaches. Table 3 shows setting parameters and performance metrics for the different models. The first column indicates the network design. The second column shows whether or not ground transportation is used. The columns *count* and *weight* indicate the number of demands served and the total weight of demands served, respectively. Column #*T* indicates the number of airports used as transshipment locations. Columns *time* and *gap* give the CPU time (in seconds) and the optimality gap, respectively. For the DS model, we report the optimality gap for the solution to the integer program with the columns generated in the column generation phase (for more details, see Yıldız and Savelsbergh (2022)).

The results clearly demonstrate the importance of considering ground transportation in the design of an effective air transportation network (the results are represented graphically in Fig. 4). Regardless of the fleet size, the demand served by weight is always significantly higher when multimodal transportation is used (results are similar to demand served by count; please see Appendix B.) However, determining a high-quality integrated (multimodal) service network is much harder computationally. When ground transportation is considered, the number of demand assignment variables increases significantly, which results in a much larger and more difficult integer program.

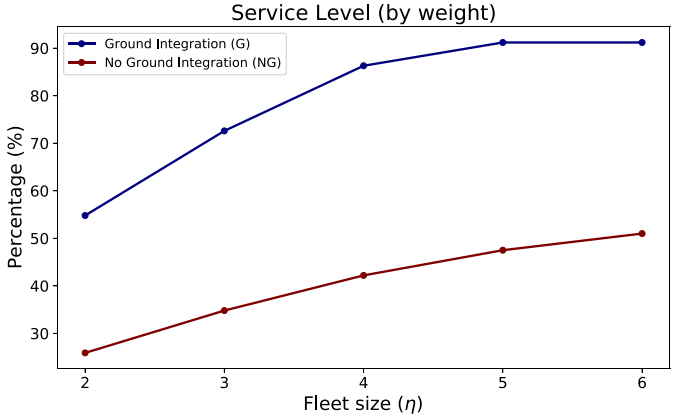

Fig. 4. Value of integrating ground transportation.

Understanding the value of transshipment requires more careful reflection. Fig. 5 illustrates the impact of fleet size on the total demand served by weight under TSM, DS, and STP. Results are similar for the total demand served by count, see Appendix C. The results show that the most effective transshipment network design always uses a few (but not that many) transshipment locations. For all instances, a transshipment network with a single transshipment location performs worse (serving up to %11.1 fewer demands by count and up to %10.6 less demand by weight).

Perhaps more interesting is the comparison with a direct shipment network. With two cargo planes, the benefits of transshipment cannot be fully realized because of the fact that each cargo plane can only perform one pickup and one delivery flight, and demands have tight time constraints. As a result, a DS design outperforms a TS design. With six cargo planes, the ground time at transshipment locations (to transfer pallets from one cargo plane to another) starts to become a limiting factor as it reduces the possible flying time. Furthermore, the advantage provided by the flexibility to load and unload at every stop in a DS design starts to show. As a result, a DS design outperforms a TS design. The results highlight a sweet spot (a sweet range more accurately)

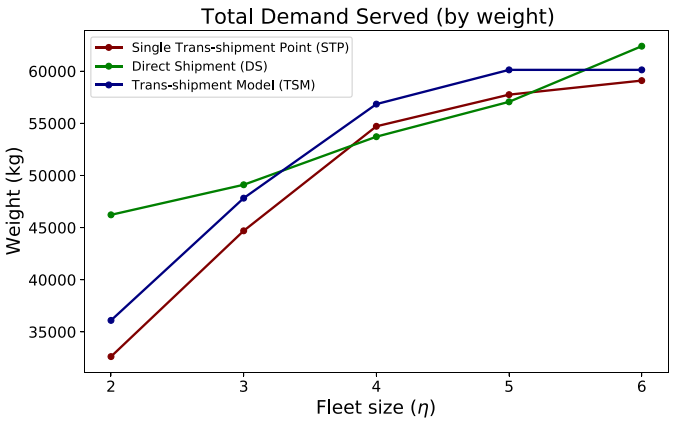

Fig. 5. Value of transshipment.

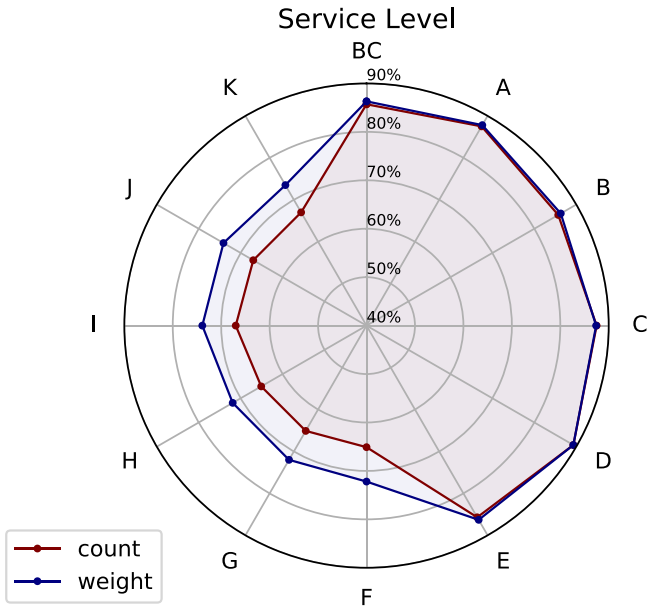

Fig. 6. The impact of operational improvements on demand served.

in terms of the number of cargo planes, where transshipment offers benefits.

The results also reveal (see Fig. 5) that increasing the number of cargo planes has diminishing returns in terms of coverage, i.e., keeping all other factors equal, the marginal contribution of an additional cargo plane decreases as the fleet size increases. As the number of cargo planes increases, it becomes more and more challenging to increase the demand served as the only unserved demands left are the "hard to reach" demands. For this set of demands, there is no benefit to increasing the fleet size beyond five cargo planes the maximum service coverage can be achieved with five cargo planes. In practice, companies may want to keep the fleet sizes even smaller. When the TS network complements a star-shaped hub & spoke (star) network, it is important to ensure that the utilization rates of the planes reserved for the TS network do not drop below the utilization rates of the planes used in the hub & spoke network. For this set of demands, reserving four planes is likely to be a good choice, as the per-plane supported demand (in terms of weight) is greater than the plane capacity (14 tons). Obviously, if resiliency is a concern, then companies may choose a slightly larger fleet to increase flexibility at the expense of increased operational costs. Therefore, the service coverage (which demands to serve) and capacity allocation (how many planes to use) decisions have to be made carefully, taking into account the distribution of demand over the service region as well as other factors such as the competitive strategy of the company (i.e., efficiency vs. agility).

Finally, We find that the number of transshipment points used does not appear to be sensitive to a change in the fleet size: for all fleet sizes, the TS network has 2–3 transshipment locations.

4.3. Impact of changes in operating environment

Available time is a critical resource in transportation network design, which is a more critical issue for the express air-cargo operators in China with limited aircraft operation times imposed by the regulations. We design and implement a set of experiments to measure the impacts of changes in the operating environment, namely aircraft operation time window, intra-city operations time, and airport operations time, using the following parameter settings:

- Cargo aircraft operation time windows: $[\underline{\phi}, \overline{\phi}] \in \{[23:00, 08:00], [23:00, 09:00], [22:00, 08:00]\}$.
- Intra-city operations time: $\phi_c \in \{4.5, 5.5\}$.
- Airport operations time: $\tau \in \{0.75, 1\}$.

Table 4 compares the performance of the 12 settings based on our key performance indicators. Columns $d.count$ and $d.weight$ refer to the number and the total weight of the available demands (that become ready in time for air cargo operations), respectively. Configuration BC represents the base case. Details of the solution for the base case (OD pairs, demand weights, ready times, repeatable pick-up and delivery paths, and TS points used) can be found in Appendix A. Configurations A, B, D, E, G, H, J, and K extend the operation time window by 1 h. Configurations A, D, G, and J finish one hour later, whereas configurations B, E, H, and K start one hour earlier. Configurations C, D, E, I, J, and K use reduced airport operation time in pickup and delivery stops (45 min). Configurations F, G, H, I, J, and K use reduced intra-city operation time (4.5 h as compared to 5.5 h in the BC). Note that, when the intra-city operations are completed faster (configurations F, G, H, I, J, and K), the number of demands that become ready for the air-cargo operations before the midnight cutoff time increases from 56 to 109, giving more opportunities to support more demand by the TS network. It is also worth noting that for these instances the transportation network consists of 26 airports and 40 cities as opposed to 24 airports and 33 cities in configurations BC, A, B, C, D, and E.

Fig. 6 reveals that improving the efficiency of intra-city operations gives the highest increase in demand served. This is mostly due to the increased number of potential demands that can be served by the TS networks, as mentioned before. Interestingly, improving the efficiency of ground operations does not have a significant impact. This finding contrasts with Yıldız and Savelsbergh (2022), the direct shipment model where the increased efficiency in ground transportation leads to a significant increase in demand satisfied. This observation highlights the different characteristics of the direct shipment and transshipment models and the demands covered by them. In the direct shipment model proposed by Yıldız and Savelsbergh (2022) to distribute express parcels, the number of flight legs in a cargo plane itinerary is of critical importance, and this number strongly depends on the time spent on ground operations. Even moderate time savings in ground operations may allow the addition of one (or more) legs to an itinerary, which in turn, can significantly increase the demand served. However, such efficiency improvements cannot easily be converted to an increase in coverage in the transshipment model due to two reasons. First, in a TS network cargo planes still need to wait for each other at the transshipment nodes (to exchange cargo), and adding one more stop to a pickup or delivery path may add delays in all connected flights. Second, a TS network already covers a large portion of the available demand in our instances, leaving little room for improvement. We also find that the timing of the flight hour extension is critical. Our results

Table 4
The impact of changes in operating environment.

ID	$\underline{\phi}$	$\overline{\phi}$	ϕ_c	τ	# airports	#cities	d.count	d.weight	Count	Weight	#T	time	Gap
BC	23	32	5.5	1	24	33	56	65,900	48	56,845	3	43,202	1.60%
A	23	33	5.5	1	24	33	56	65,900	49	57,874	3	43,202	3.90%
B	22	32	5.5	1	24	33	56	65,900	48	56,845	3	43,203	3.85%
C	23	32	5.5	0.75	24	33	56	65,900	49	57,629	3	43,205	4.34%
D	23	33	5.5	0.75	24	33	56	65,900	50	58,879	2	43,205	2.12%
E	22	32	5.5	0.75	24	33	56	65,900	48	56,845	3	43,206	5.78%
F	23	32	4.5	1	26	40	109	119,923	71	86,588	3	43,204	3.89%
G	23	33	4.5	1	26	40	109	119,923	71	86,388	2	43,203	4.68%
H	22	32	4.5	1	26	40	109	119,923	71	86,236	3	43,204	4.28%
I	23	32	4.5	0.75	26	40	109	119,923	73	88,623	3	43,210	4.04%
J	23	33	4.5	0.75	26	40	109	119,923	73	88,899	2	43,207	3.63%
K	22	32	4.5	0.75	26	40	109	119,923	73	88,117	3	43,207	4.55%

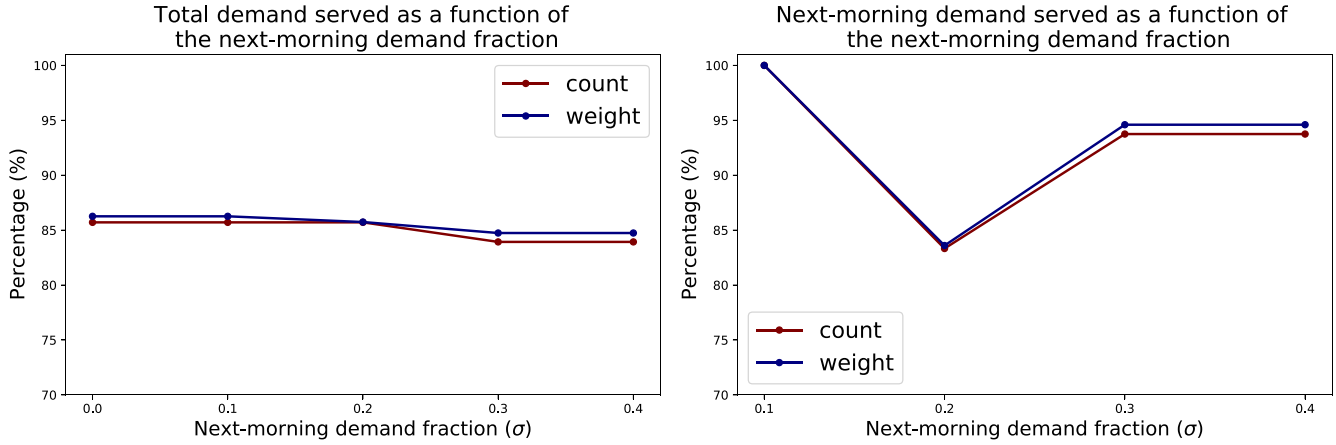

Fig. 7. Impact of service class distribution on the service level.

Table 5
The impact of service class distribution.

σ	Count	Weight	#T	Time	Gap
0	48	56,845	3	43,202	1.6%
0.1	48	56,845	3	43,202	3.80%
0.2	48	56,502	2	43,203	2.21%
0.3	47	55,843	3	43,202	4.21%
0.4	47	55,843	3	43,203	5.60%

show that starting the cargo flights one hour earlier (configurations B, E, H, and K) does not make any difference in the amount of demands served. Interestingly, starting one hour earlier leads to a slightly worse performance (ranging between −0.4% and −1.4%) in terms of the total weight of demand served for configurations E, H, and K, where either intra-city operations time, airport operations time, or both shorten. However, extending the flight time by one hour has a significant impact, particularly for configurations A and D, where service level by count and weight increase by about 2%.

4.4. Impact of service class distribution

To capture the impact of offering different service class compositions, we experiment on four configurations: $\sigma = \{0, 10\%, 20\%, 30\%, 40\%\}$, where σ represents the percentage of shipments that need to be delivered next morning before noon (i.e., as σ increases, the proportion of critical items with time priority increases).

Table 5 summarizes the results of our computational experiments. Fig. 7 illustrates the impact of service class distribution on the total demand and the next morning's demand served. The left and right figures depict the total demand, and the next morning served by value and count, respectively. We see that the impact of demand composition does not have a significant impact on almost any of our performance metrics. These interesting results show that for the problem instances we study in this paper, demands that could be served with the next-day delivery deadline can also be served. This result verifies that by shortening the travel distances (times) and by eliminating the time needed for the sorting operations a TS network can complete cargo transfers fast enough to meet demanding delivery deadlines.

5. Conclusion

Given the push towards shorter and shorter lead times, the need for more agile supply chains, as well as other demand and supply dynamics, experts predict that the market for express shipment services will continue to grow worldwide in the next two decades.

In this study, we propose a novel transshipment-based express shipment network design and develop an optimization model to identify the design that maximizes the demand served (in terms of total weight). The compact path-based mixed-integer program, incorporating various spatial and temporal constraints, allows solving practical instances in a reasonable amount of time.

An extensive computational study reveals important managerial insight. First, the integration of air and ground transportation considerably improves the demand coverage. Second, the use of transshipment enables serving a high number of origin–destination pairs with just a small number of cargo planes while meeting tight lead time requirements. These two findings indicate that express air cargo companies do not necessarily have to invest vast sums of money in additional capacity to be able to offer fast service across a large area. Careful planning and coordination of ground and air transportation capacities and exploiting transshipment may provide a viable alternative. Finally, analysis reveals that shortening intra-city operations and extending the nightly operating period are the most effective adjustments that can increase the performance of the proposed transshipment network.

Our study is only a first step towards developing a hybrid design incorporating direct shipment, transshipment, and (traditional) star-shaped hub & spoke networks to meet diversifying customer needs in air cargo service. Further research is needed to explore how to construct high-quality hybrid designs and best operate hybrid designs (e.g., the allocation of express cargo to the different networks). Other relevant research could explore how to manage the service network over time given exogenous factors, such as the inherent seasonality in demand at the presence of tight regulations that allow only gradual changes in the flight schedules. As cargo flight plans must be approved by authorities and cannot be changed easily, developing robust designs against uncertain changes in demand is yet another avenue for future research. As is the case for many network design problems, the scalability of the solution approach is a concern. As the instance size increases, the computational complexity increases exponentially. Although we were able to solve realistic-size instances arising in our case study for SF Express, new ideas/techniques may be needed to improve the computational efficiency to solve larger instances that may arise in different settings.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful for the generous financial support by the Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 2219 and thankful for the data and assistance provided by the air service network planning team at SF Express. This paper would not be possible without the valuable input of these organizations.

Appendix A. Solution for the base case

Table A.6 presents our solution for the Base Case. The column "ID" indicates the unique demand identifier for a demand $q \in Q$ with origin o^q, destination d^q, deadline t^q, and weight f^q. Demand coverage by the TS network are shown in the column "Covered?", where 1 indicates that the demand is covered and 0 shows it is not covered. The gateway airports where the demand enters and leaves the air transport network are shown in the columns "Picked-up" and "Dropped." As their name suggest, the columns "Pick-up Path" and "Delivery Path" show the pickup and delivery paths used by the respective demand (including the empty paths marked as such). The column "TS Node" shows the transshipment node for the covered demands, and the gateway airport where the demand leaves the air transport network is indicated in the column "Delivery path."

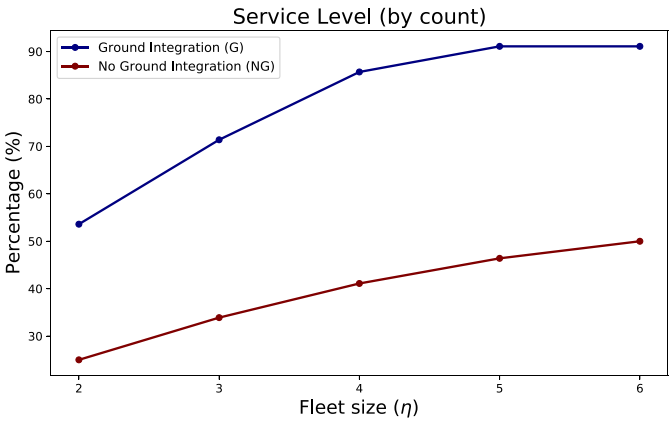

Fig. 8. Value of integrating ground transportation.

Appendix B. Value of ground integration

See Fig. 8.

Appendix C. Value of transshipment

See Fig. 9.

Table A.6
Solution for the base case.

ID	o^q	d^q	t^q	f^q	Covered?	Picked-up	Pick-up path	TS node	Delivery path	Dropped
1	1	115	23	1002	1	1	[1- 144- 3]	3	[3- 112- 38]	112
2	1	138	15	1208	1	1	[1- 144- 3]	3	Empty	3
4	2	3	15	1250	1	2	[8- 2- 138]	138	Empty	138
5	2	149	23	1250	1	2	[8- 2- 138]	138	[2- 210- 152]	152
6	2	152	23	1250	1	2	Empty	2	[2- 210- 152]	152
7	2	210	23	1250	1	2	Empty	2	[2- 210- 152]	210
8	2	318	23	1250	1	2	[8- 2- 138]	138	[138- 318- 8]	318
9	3	1	15	1250	1	3	[38- 116- 3]	3	[3- 123- 1]	1
10	3	1	19	1250	1	3	Empty	3	[3- 123- 1]	1
11	3	4	21	1206	1	3	Empty	3	[3- 123- 1]	1
13	3	8	21	1250	1	138	Empty	138	[138- 318- 8]	8
14	3	112	21	1250	1	138	[8- 2- 138]	138	[3- 112- 38]	112
15	3	115	23	1250	1	3	Empty	3	[3- 112- 38]	112
16	3	116	23	1250	1	3	Empty	3	[3- 112- 38]	112
17	3	130	23	1250	1	3	Empty	3	[3- 123- 1]	123
18	3	135	23	1051	1	3	Empty	3	[3- 123- 1]	123
20	6	1	19	1020	0	–	–	–	–	–
22	7	8	23	1017	1	138	Empty	138	[138- 318- 8]	8
24	7	144	23	1005	0	–	–	–	–	–
26	8	3	23	1250	1	8	[8- 2- 138]	138	Empty	138
27	10	9	23	1064	0	–	–	–	–	–

(*continued on next page*)

Table A.6 (*continued*).

ID	o^q	d^q	t^q	f^q	Covered?	Picked-up	Pick-up path	TS node	Delivery path	Dropped
29	10	275	23	1187	0	–	–	–	–	–
31	38	1	23	1250	1	38	[38- 116- 3]	3	[3- 123- 1]	1
32	54	1	15	1250	0	–	–	–	–	–
33	98	1	21	1250	1	3	Empty	3	[3- 123- 1]	1
34	100	4	21	1250	1	3	Empty	3	[3- 123- 1]	1
35	100	8	21	1250	1	138	[8- 2- 138]	138	[138- 318- 8]	8
36	100	8	23	1250	1	138	Empty	138	[138- 318- 8]	8
37	100	38	13	1029	1	3	Empty	3	[3- 112- 38]	38
39	100	38	21	1089	1	3	Empty	3	[3- 112- 38]	38
41	100	38	23	1023	1	3	Empty	3	[3- 112- 38]	38
43	100	112	21	1065	1	3	[38- 116- 3]	3	[3- 112- 38]	112
45	100	112	23	1250	1	3	Empty	3	[3- 112- 38]	112
46	100	200	13	1250	1	138	[152- 138- 2]	2	Empty	2
47	100	208	13	1250	1	138	[152- 138- 2]	2	Empty	2
48	100	208	15	1250	1	138	[152- 138- 2]	2	Empty	2
49	100	318	13	1155	1	138	Empty	138	[138- 318- 8]	318
51	100	318	21	1063	1	138	Empty	138	[138- 318- 8]	318
53	104	3	21	1250	0	–	–	–	–	–
54	111	3	23	1250	1	116	[38- 116- 3]	3	Empty	3
55	112	1	21	1169	1	116	[38- 116- 3]	3	[3- 123- 1]	1
57	112	3	21	1236	1	116	[38- 116- 3]	3	Empty	3
59	115	1	23	1029	0	–	–	–	–	–
61	141	123	23	1080	1	3	[38- 116- 3]	3	[3- 123- 1]	123
63	144	7	23	1250	1	144	[1- 144- 3]	3	[3- 123- 1]	123
64	149	2	23	1250	0	–	–	–	–	–
65	149	198	21	1020	1	152	[152- 138- 2]	2	Empty	2
67	152	2	23	1250	1	152	[152- 138- 2]	2	Empty	2
68	152	198	23	1223	1	152	[152- 138- 2]	2	Empty	2
70	152	208	23	1001	1	152	[152- 138- 2]	2	Empty	2
72	198	3	13	1075	1	2	[8- 2- 138]	138	Empty	138
74	198	148	23	1250	1	2	[8- 2- 138]	138	[2- 210- 152]	152
75	198	149	23	1250	1	2	[8- 2- 138]	138	[2- 210- 152]	152
76	198	152	23	1177	1	2	Empty	2	[2- 210- 152]	152
78	198	210	23	1127	1	2	Empty	2	[2- 210- 152]	210
80	198	318	23	1079	1	2	[8- 2- 138]	138	[138- 318- 8]	318

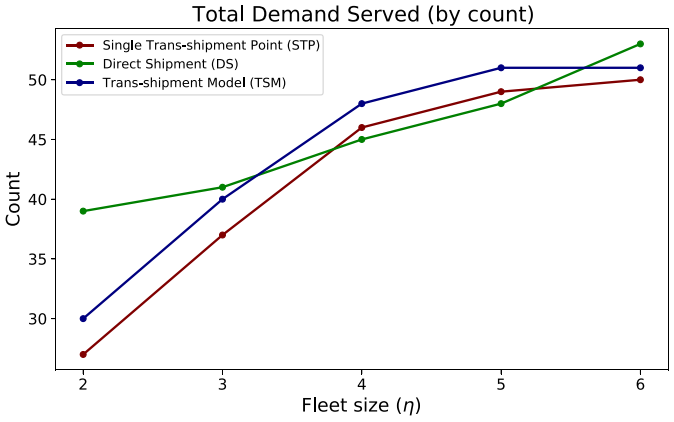

Fig. 9. Value of transshipment.

References

Agarwal, Y.K., 2002. Design of capacitated multicommodity networks with multiple facilities. Oper. Res. 50 (2), 333–344.

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 J. Oper. Res. 291 (1), 1–17.

Alumur, S.A., Kara, B.Y., Karasan, O.E., 2012a. Multimodal hub location and hub network design. Omega 40 (6), 927–939.

Alumur, S.A., Yaman, H., Kara, B.Y., 2012b. Hierarchical multimodal hub location problem with time-definite deliveries. Transp. Res. E 48 (6), 1107–1120.

Armacost, A.P., Barnhart, C., Ware, K.A., 2002. Composite variable formulations for express shipment service network design. Transp. Sci. 36 (1), 1–20.

Barnhart, C., Krishnan, N., Kim, D., Ware, K., 2002. Network design for express shipment delivery. Comput. Optim. Appl. 21 (3), 239–262.

Büdenbender, K., Grünert, T., Sebastian, H.-J., 2000. A hybrid tabu search/branch-and-bound algorithm for the direct flight network design problem. Transp. Sci. 34 (4), 364–380.

Cohn, A.M., Barnhart, C., 2003. Improving crew scheduling by incorporating key maintenance routing decisions. Oper. Res. 51 (3), 387–396.

Crabtree, T., Hoang, T., Gildemann, G., Gildemann, G., 2020. World air cargo forecast. https://www.boeing.com/commercial/market/cargo-forecast/. (Accessed 09 February 2022).

Crainic, T.G., 2000. Service network design in freight transportation. European J. Oper. Res. 122 (2), 272–288.

Dai, W., Wandelt, S., Zhang, J., Sun, X., 2021. Capacitated air/rail hub location problem with uncertainty: A model, efficient solution algorithm, and case study. IEEE Trans. Intell. Transp. Syst. 23 (7), 8451–8466.

Derigs, U., Friederichs, S., 2013. Air cargo scheduling: integrated models and solution procedures. OR Spectrum 35 (2), 325–362.

Derigs, U., Friederichs, S., Schäfer, S., 2009. A new approach for air cargo network planning. Transp. Sci. 43 (3), 370–380.

Feng, B., Li, Y., Shen, Z.-J.M., 2015. Air cargo operations: Literature review and comparison with practices. Transp. Res. C 56, 263–280.

IATA, 2021. Annual review. https://www.iata.org/en/publications/annual-review/. (Accessed 09 February 2022).

Kim, D., Barnhart, C., Ware, K., Reinhardt, G., 1999. Multimodal express package delivery: A service network design application. Transp. Sci. 33 (4), 391–407.

Lee, C.K., Zhang, S., Ng, K.K., 2019. Design of an integration model for air cargo transportation network design and flight route selection. Sustainability 11 (19), 5197.

Lin, C.-C., Chen, Y.-C., 2003. The integration of Taiwanese and Chinese air networks for direct air cargo services. Transp. Res A 37 (7), 629–647.

Lin, C.-C., Chen, S.-H., 2008. An integral constrained generalized hub-and-spoke network design problem. Transp. Res. E 44 (6), 986–1003.

Louwerse, I., Mijnarends, J., Meuffels, I., Huisman, D., Fleuren, H., 2014. Scheduling movements in the network of an express service provider. Flex. Serv. Manuf. J. 26 (4), 565–584.

Prodhon, C., Prins, C., 2014. A survey of recent research on location-routing problems. European J. Oper. Res. 238 (1), 1–17.

Qu, Y., Bard, J.F., 2012. A GRASP with adaptive large neighborhood search for pickup and delivery problems with transshipment. Comput. Oper. Res. 39 (10), 2439–2456.

Serper, E.Z., Alumur, S.A., 2016. The design of capacitated intermodal hub networks with different vehicle types. Transp. Res. B 86, 51–65.

Smilowitz, K.R., Daganzo, C.F., 2007. Continuum approximation techniques for the design of integrated package distribution systems. Netw. Int. J. 50 (3), 183–196.

SteadieSeifi, M., Dellaert, N.P., Nuijten, W., Van Woensel, T., Raoufi, R., 2014. Multimodal freight transportation planning: A literature review. European J. Oper. Res. 233 (1), 1–15.

Wandelt, S., Dai, W., Zhang, J., Sun, X., 2022. Toward a reference experimental benchmark for solving hub location problems. Transp. Sci. 56 (2), 543–564.
Yaman, H., 2009. The hierarchical hub median problem with single assignment. Transp. Res. B 43 (6), 643–658.
Yan, S., Chen, C.-H., 2008. Optimal flight scheduling models for cargo airlines under alliances. J. Sched. 11 (3), 175–186.
Yan, S., Chen, S.-C., Chen, C.-H., 2006. Air cargo fleet routing and timetable setting with multiple on-time demands. Transp. Res. E 42 (5), 409–430.
Yıldız, B., Savelsbergh, M., 2022. Optimizing package express operations in China. European J. Oper. Res. 300 (1), 320–335.
Yıldız, B., Yaman, H., Karaşan, O.E., 2021. Hub location, routing, and route dimensioning: Strategic and tactical intermodal transportation hub network design. Transp. Sci. 55 (6), 1351–1369.
Yu, S., Yang, Z., Yu, B., 2017. Air express network design based on express path choices–Chinese case study. J. Air Transp. Manag. 61, 73–80.