Journal of Air Transport Management 73 (2018) 95–112

Contents lists available at ScienceDirect

Journal of Air Transport Management

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

The gateway hub location problem

Luiza Bernardes Real[a,*], Morton O'Kelly[b], Gilberto de Miranda[c], Ricardo Saraiva de Camargo[d]

[a] Instituto Federal de Minas Gerais, Brazil
[b] Department of Geography, Ohio State University, OH, USA
[c] Departamento de Matem´atica Aplicada, Universidade Federal do Esprito Santo, Brazil
[d] Departamento de Engenharia de Produ¸cao, Universidade Federal de Minas Gerais, Brazil

ARTICLE INFO

Keywords:
Air transport
Hub-and-spoke networks
Gateway
Benders decomposition method

ABSTRACT

We introduce the Gateway Hub Location Problem (GHLP) to design global air transportation systems. Relying on a three-level hub network structure and on having nodes located in different geographic regions, the GHLP consists of locating international gateways and domestic hubs, activating arcs to induce a connected gateway and hub network, and routing flows within the network at minimum cost. Most previous studies focus on a typical hub-and-spoke network, in which local and global flows are not differentiated. Here to better represent a world wide air transportation system, global flows can only leave or enter a given geographic region by means of a gateway, while local flows can only use hubs within their respective region. As routing local or global flows involved different agents, this study presents a mixed integer programming formulation that exploits these differences to model both the local and global flows. Due to the formulation's characteristics, two algorithm variants based on Benders decomposition method are devised to solve the problem. A new repair procedure produces optimality Benders cuts whenever feasibility Benders cuts would rather be expected. While the monolithic version failed to solve medium size instances, our algorithms solved lager ones in reasonable time.

1. Introduction

By the year 2034, global air traffic is expected to double reaching over seven billion passengers annually transported (IATA, 2015), being Africa, Middle East, Asian and Latin America the geographic areas with the largest percentage growths till then. This rapid demand growth is pressuring airlines and air transport management agencies to extend and expand the existent networks to accommodate new markets, new players, new infra-structures, and new flight connections to serve both increasing domestic (local) and international (global) passenger flows.

Modeling and understanding these local and global passenger flows are generally done separately in the literature (Preis et al., 2013; Mao et al., 2015), or are usually considered to be non differentiable when designing networks for many-to-many air transportation systems with a hub-and-spoke structure (Campbell et al., 2002; Alumur and Kara, 2008; Campbell and O'Kelly, 2012; Farahani et al., 2013). However, there are some differences between domestic and international passengers that might justify differentiating them.

From the perspective of service quality, reliability was ranked by international passengers as the most important dimension, whereas domestic passengers value more assurance dimension (Arslan et al., 2011). According to the Resource Manual for Airport In-Terminal Concessions (2011), international passengers, on average, arrive at the airport earlier and spend more time in terminals. Thereby, their needs for food, reading materials, travel accessories and other amenities are larger than domestic passengers' needs. They also tend to be more sophisticated with higher average incomes. This represents a higher potential revenue for international airports, which usually have a better infrastructure for shopping, eating and even resting when contrasting to domestic airports. In this way, even though they might share some resources, and affect each other's routing design decisions, when designing air passenger networks, local and global flows are required to be routed through different facility types over the network.

Local flows are routed via domestic hubs (hubs), while global flows go through international gateways (gateways) to leave from or to enter into a different geographic region. Hubs allow passengers to change connections and airplanes along their routes, whereas gateways are critical for connecting wide regions, such as continents, and for performing customs, immigration and security checks. Since a global passenger flow may be routed via some hubs before going through some

* Corresponding author. Departmento de Engenharia de Producao, Universidade Federal de Minas Gerais, Av. Antonio Carlos, 6627-31270-901, Belo Horizonte, MG, Brazil.

E-mail addresses: luiza.real@ifmg.edu.br (L. Bernardes Real), okelly.1@osu.edu (M. O'Kelly), gilberto.miranda@ufes.br (G. de Miranda), rcamargo@dep.ufmg.br (R. Saraiva de Camargo).

https://doi.org/10.1016/j.jairtraman.2018.08.006
Received 30 June 2017; Received in revised form 16 May 2018; Accepted 20 August 2018
0969-6997/ © 2018 Elsevier Ltd. All rights reserved.

gateways to reach its destination, or vice versa, a global flow may share then some inter-hub connections with other local flows, showing thus how both flow types are intertwined.

To articulate both local and global flows, three connection levels are needed: international gateway level with inter-gateway connections, domestic hub level with inter-hub connections, and spoke level with regional airports linked to hubs or gateways. Inter-hub connections are usually done by large carriers, while inter-gateway connections are performed by even larger, long-range airplanes. Further regional airports are usually linked to hubs or gateways by middle to small size planes. This three level setting can be seen as a three-tier hierarchical hub-and-spoke network structure.

Hub-and-spoke systems are commonly used in many-to-many transport applications to lower transportation costs by exploiting scale economies whenever large carriers can be used to carry consolidated flows over the network (O'Kelly, 1987; Jr, 2012). A typical hub-and-spoke network uses two connection levels instead of three: hub level with inter-hub connections, and spoke level with flow exchanging nodes (regional airports) linked to hubs. Scale economies are usually achieved on the hub level by bulk transportation on inter-hub connections. A myriad of applications and topologies have been modeled as hub-and-spoke networks as can be seen in Campbell et al. (2002); Alumur and Kara (2008); Campbell and O'Kelly (2012); Farahani et al. (2013).

In the past 20 years the global airline industry has undergone major changes. The notion of international airlines collaborating for creating cost and revenue synergies through the formation of strategic alliances (such as Star Alliance, Oneworld and Skyteam) has been gained credibility (Schosser and Wittmer, 2015). Thereby, design air network from a global perspective becomes necessary. However, only in the last decade, the idea of differentiating local from global flows has attracted some attention from the research community (Adler and Smilowitz, 2007; Sasaki et al., 2009; Yaman, 2009; Catanzaro et al., 2011).

Adler and Smilowitz (2007) analyze global alliances and mergers in an airline industry under competition. They present a game-theoretic competitive merger framework that allows airlines to choose partners with their installed gateways, inter-gateway connections, and regional networks so that mergers can be proposed and profits maximized. Selection is based on cost and revenue analyses by considering information of a given airlines and its competitors. Local and global flows are differentiated, but treated separately on a two stage approach. First hubs are installed to route local flows, then, assuming that global flows are temporarily aggregated at each installed hub, rather than in their original locations, one gateway per region is selected within these installed hubs. As the trace of each demand flow exchange can only be performed after the network is designed, transportation costs are poorly underestimated, questioning thus the quality of the achieved network configurations.

Disregarding the many-to-many nature of the local and global flows, Sasaki et al. (2009) develop a gateway and hub location model based on a two level *p*-median facility location problem. From a candidate set, a fixed number of gateways and hubs are selected so that each regional airport is served by a hub, and each installed hub is linked to a gateway at minimum allocation cost. By not considering flow demands happening between pairs of origin-destination nodes, the problem's complexity is greatly reduced at the expense of having ill-formed air networks.

Yaman (2009) does not distinguish between local and global flows, he considers the design of a hierarchical hub and spoke network which consists of locating a fixed number of gateways and hubs, such that regional airports and hubs are single allocated to hubs and gateways, respectively, to form a star sub-network for each gateway. The simpler strict formulation imposes gateways to be fully interconnected, and prevent hubs to directly interact with each other. Given the single allocation policy, undesirable long distances are perceived by the demand flows in the attained solutions. Further, a fully interconnected gateway hub is not always possible to be assumed in an air network design, since airlines tend to avoid flying for long ranges over water without communication, or over conflict zones.

Finally, Catanzaro et al. (2011) investigate a particular variant of a hub location problem which partitions a given network into sub-networks, and locates at most a fixed number of gateways, but with at least one gateway in each sub-network. Sub-networks are supposed to have at least (at most) a minimum (maximum) number of nodes to exist. The problem's objective is to split the network into regions and then route flows at minimum transportation cost. A flow can only enter or leave a sub-network through an installed gateway, and once it leaves a sub-network, it can only be routed through gateways until it reaches its destination sub-network, when then it can use the available hubs and local links. Hubs and all network connections are assumed to be given beforehand, i.e. costs incurred from installing hubs, gateways, and inter-hub and inter-gateway connections are not considered.

Until now, the literature has acknowledged the importance of differentiating local from global flows, but, as aforementioned, has made assumption compromises that resulted into over-simplified problems or models. Here a more explicit formulation that incorporates local and global flows is proposed for the air transportation network design. Hubs, gateways, and inter-hub and inter-gateway connections are decided so that the induced network can route local and global flows at minimal transportation and installation costs. Different scale economies are granted for installed inter-hub and inter-gateway connections to mimic lower transportation costs due to consolidated flows. Regional airports can be linked to any installed hub or gateway within its region and within aircraft range, i.e. local airports can be multiple allocated to hubs and gateways. This provides greater flexibility to route flows at the expense of demanding a more elaborated model. Further fixed costs for establishing hubs, gateways, and inter-hub and inter-gateway connections are assumed to be known, and continental and country divisions are adopted as natural regions.

Because our aim is to consider the design of air network from a global perspective, we made some simplifications for now. The current study ignored, for example: airline competition, passengers behavior and choice of routes, congestion transshipment airports, the effects of frequency on service quality and schedule delay. These issues have been well studied in the literature (Hansen, 1990; Hong and Harker, 1992; Hsu and Wen, 2003; Adler, 2005).

The addressed air transportation network design is modeled as a multi-commodity flow based hub and spoke system, given rise to a gateway hub location problem or a three-level hub location problem. Given its large scale multi-commodity nature and its induced decomposable matrix structure, the devised formulation is solved by two specialized Benders decomposition algorithms (Benders, 1962) which incorporate two features that greatly speed up the method: a repair procedure which allows to generate Benders optimality cuts from unbounded dual subproblems, and a tailored dual subproblem solution algorithm which calculates the optimal dual values to produce Benders optimality cuts without relying on a Simplex solver. In order to evaluate and assess the efficiency and limitations of the devised Benders algorithms, computational experiments were performed and compared with a general purpose solver (IBM CPLEX) on solving the proposed formulation. Both algorithms clearly out-performed the general purpose solver when solving large instance sizes.

To be clear from the outset, the focus here is not on reproducing the current air network, rather, we wish to use network design tools that contribute to improve air transport systems. The proposed model of how things should be can be used to contrast to actual systems. We believe that this analysis is needed and should be of concern. There are many broad participants interested in the efficiency of the world's aviation system as World Bank, FAA (Federal Aviation Administration), Eurocontrol, mainframe manufacturers (Boeing, Airbus, Embraer) and probably many others. The rational planning of the air network has implications consistent with the strategic objectives of ICAO

(International Civil Aviation Organization) on supporting the growth of air transport. Analyses of the efficient of aviation could or ought to be used to harmonize the air transport framework focused on the development of an economically viable aviation infrastructure. This study suggests a move towards a more rational conception of air network design, based on differentiating local and global air passenger flows.

Operations research (OR) has played a critical role in helping the airline industry designs its air network, plans its scheduling, routing and crew assignment (Barnhart et al., 2003). Since 1961, there is a professional society dedicated to the advancement and application of Operational Research within the airline industry, the Airline Group of the International Federation of Operational Research Societies (AGIFORS). More than 500 airlines and air transport associations are currently represented in AGIFORS.

The contribution of this study is twofold: To propose a more explicit formulation that incorporates local and global flows to design air transportation networks with a hub and spoke structure, and to devise exact algorithms capable of solving large-scale instances in reasonable time. The remainder of this article is organized as follows. §2 introduces the used notation and the proposed mathematical formulation; while §3 presents the devised Benders algorithms to solve the problem. §§4 and 5 reports the carried out computational experiments, and the insights provided by this study, respectively. §6 discusses achieved conclusions.

2. Notation, definitions and formulation

Given different geographic regions with airports which exchange flows between them, the addressed gateway hub location problem consists of locating hubs and gateways, and inter-hub and inter-gateway connections so that a hub and spoke based air network is designed, and local and global flows can be routed at minimal transportation and fixed costs. Local flows are not required to go through hubs to be routed, i.e. direct connections between local airports are allowed. Fixed costs for direct connections are disregarded, but transportation costs are accounted for them. However scale economies are only granted on inter-hub and inter-gateway connections which are required to be established. Global flows or flow exchanges between airports of different regions, pass through at least two gateways: one in the origin's region, and another in the destination's region. Consequently, each region must have at least one gateway. A gateway also acts as a hub, but a hub does not operate as a gateway unless one is installed at it. Local airports can be connected to more than one hub or gateway within their region, in other words, multiple allocations are allowed. Further the gateway level form a connected incomplete network.

To model the aforementioned problem, three different layers can be used to represent each airport operation type: layer *I* has all the actual local airports of each region, layer *II* contains hub candidate airports, while layer *III* has gateway candidate airports. To illustrate this idea, please refer to Fig. 1. The original configuration represented by Fig. 1a has three different regions with six local airports, which compose layer *I*. Assuming that each local airport is a hub and a gateway candidate for this example, then, by copying each local airport of layer *I* and re-indexing and relabeling it accordingly, it is possible to set layers *II* (airports $7-12$) and *III* (airports $13-18$). Please see Figure b.

These three layers can be represented by a digraph $\mathbb{G} = (N, A)$ in which N and A are the airport and the arc sets, respectively. Set $N = L \cup H \cup G$ is formed by disjoint subsets L (local airport set), H (hub candidate airport set), and G (gateway candidate airport set)– note that set N has its airports re-indexed as aforementioned – while set A has to be carefully assembled. Let $\mathscr{R} = \{1, \ldots, n_{\mathscr{R}}\}$ be the region set, in which $n_{\mathscr{R}}$ is the number of regions, and $\phi(i)$ and $R(i)$ are two mathematical functions that return the original airport index and the region of airport $i \in N$, respectively. In the example of Fig. 1, $\mathscr{R} = \{1,2,3\}$, $L = \{1, \ldots, 6\}$, $H = \{7, \ldots, 12\}$, and $G = \{13, \ldots, 18\}$, and as e.g. $\phi(2) = \phi(8) = \phi(14) = 2$ and $R(6) = R(12) = R(18) = 3$. Set A is then composed of different arc sets or $A = A^L \cup A^H \cup A^G \cup A^{LH} \cup A^{HG}$, in which $A^L = \{(i, j): i, j \in L \wedge R(i) = R(j) \wedge i \neq j\}$ is the local arc set, $A^H = \{(i, j): i, j \in H \wedge R(i) = R(j) \wedge i \neq j\}$ is the hub arc set, $A^G = \{(i, j): i, j \in G \wedge i \neq j\}$ is the gateway arc set. Note that gateway arcs can connect gateways within the same region or from different regions. Sets $A^{LH} = \{(i, j): (i \in L \wedge j \in H) \vee (i \in H \wedge j \in L) \wedge \phi(i) = \phi(j)\}$ and $A^{HG} = \{(i, j): (i \in H \wedge j \in G) \vee (i \in G \wedge j \in H) \wedge \phi(i) = \phi(j)\}$ have arcs connecting local airports to their associated hub candidates, and likewise arcs linking hub candidates to their associated gateway candidates, respectively.

Further definitions are required to model the problem. Let $W = W^l \cup W^g$ be the demand set with pairs of airports (i, j) exchanging w_{ij} units of flow, and that is formed by two subsets. $W^l = \{(i, j): i, j \in L \wedge R(i) = R(j) \wedge w_{ij} > 0\}$ and $W^g = \{(i, j): i, j \in L \wedge R(i) \neq R(j) \wedge w_{ij} > 0\}$ are sets with pairs of airports which exchange flows on a local and global level, respectively. Let also c_{uv} be a non-negative unitary cost of arc $(u, v) \in A$ given as:

$$c_{uv} = \begin{cases} \tilde{c}_{uv} \ \forall \ (u, v) \in A^L \\ \alpha^H \tilde{c}_{uv} \ \forall \ (u, v) \in A^H \\ \alpha^G \tilde{c}_{uv} \ \forall \ (u, v) \in A^G \\ b^H \ \forall \ (u, v) \in A^{LH} \\ b^G \ \forall \ (u, v) \in A^{HG} \end{cases}$$

in which $\tilde{c}_{uv}$ is the unitary transportation cost of arc (u, v), and $0 < \alpha^H < 1$ and $0 < \alpha^G < 1$ are granted discount factors to represent local and global scale economies, respectively. Parameters b^H and b^G are unitary operational costs due to baggage handling and custom and immigration checks done at hub and gateway levels, respectively.

To simplify notation, let $E = E^H \cup E^G$ be the edge set in which $E^H = \{(i, j) \in A^H: i < j\}$ and $E^G = \{(i, j) \in A^G: i < j\}$ are the edge sets associated to hub and gateway levels. Whenever an edge $(u, v) \in E$ is established to allow or inter-hub or inter-gateway flows a fixed cost q_{uv} given as:

$$q_{uv} = \begin{cases} \ell^H \tilde{q}_{uv} \ \forall \ (u, v) \in E^H \\ \ell^G \tilde{q}_{uv} \ \forall \ (u, v) \in E^G \end{cases}$$

is incurred, in which $\tilde{q}_{uv}$ is the fixed cost of edge (u, v), and ℓ^H and ℓ^G are fixed scaling factors for setting inter-hub and inter-gateway connections, respectively. Usually $\ell^H < \ell^G$. A fixed cost a_u is set for installing a hub or a gateway at airport $u \in H \cup G$.

With the aforementioned notation and definitions, and with the implementation of the layers, it is now possible to model the problem as multi-commodity flow based formulation with the help of the following variables: Let $y_u \in \{0,1\}$ be equal to 1, if a hub or a gateway $u \in H \cup G$ is installed, 0 otherwise. Further let $x_{u,v} \in \{0,1\}$ be equal to 1, if edge $(u, v) \in E$ is activated, 0, otherwise, and let $f_{uv}^{ij} \geq 0$ represent the flow percentage of demand w_{ij}, $(i, j) \in W$, that goes through arc $(u, v) \in A$. The formulation for the gateway hub location problem can now be written as:

$$\min \sum_{u \in H \cup G} a_u y_u + \sum_{(u,v) \in E} q_{uv} x_{uv} + \sum_{(i,j) \in W} \sum_{(u,v) \in A} w_{ij} c_{uv} f_{uv}^{ij} \tag{1}$$

$$\text{s.t.}: \sum_{(i,v) \in A} f_{iv}^{ij} = 1 \qquad \forall \ (i, j) \in W \tag{2}$$

$$-\sum_{\substack{(v,u) \in A \\ j \neq v}} f_{vu}^{ij} + \sum_{\substack{(u,v) \in A \\ i \neq v}} f_{uv}^{ij} = 0 \qquad \forall \ (i, j) \in W, u \in N: i \neq u, j \neq u \tag{3}$$

$$-\sum_{(u,j) \in A} f_{uj}^{ij} = -1 \qquad \forall \ (i, j) \in W \tag{4}$$

$$f_{uv}^{ij} + f_{vu}^{ij} \leq x_{uv} \qquad \forall \ (i, j) \in W, (u, v) \in E \tag{5}$$

$$\sum_{(u,v) \in A} f_{uv}^{ij} \leq y_u \qquad \forall \ (i, j) \in W, u \in H \cup G \tag{6}$$

$$\sum_{\substack{(u,v)\in E^G \\ u\in S \\ v\in G\setminus S}} x_{uv} + \sum_{\substack{(v,u)\in E^G \\ u\in G\setminus S \\ v\in S}} x_{vu} \geq y_k + y_m - 1 \qquad \forall\, k \in G\setminus S,\, m \in S\colon S\subset \mathbb{S} \wedge R(k) \neq R(m) \tag{7}$$

$$y_u \in \{0,1\} \qquad \forall\, u \in H\cup G \tag{8}$$

$$x_{u,v} \in \{0,1\} \qquad \forall\, (u,v)\in E \tag{9}$$

$$f_{uv}^{ij} \geq 0 \qquad \forall\, (i,j)\in W, (u,v)\in A \tag{10}$$

in which $\mathbb{S} = \{S\colon S\subset G, |S| \geq 2\}$.

The objective function (1) minimizes the total cost consisted of the transportation costs, and the fixed installation costs for hubs and gateways, and for inter-hub and inter-gateway connections. Constraints (2)–(4) are the flow balancing equations. The constraints (5) guarantee that flows only go through inter-hub and inter-gateway connections if their respective edges are set. Constraints (6) ensure that flows can only go through a hub or a gateway $u \in H\cup G$, if u is set. Constraints (7) are the well-known sub-tour elimination constraints (SECs) which ensure that the gateway level is always connected. They can be disregarded whenever there are global flows from at least one region to all the other regions, otherwise they are required to ensure connectivity on the gateway level. Constraints (8)–(10) set the variables' domain.

The matrix associated with the constraints' set of formulation (1)–(10) has a stair-case shape regarding the large scale variables f, but which is coupled by the integer variables y and x. This feature makes the whole system amenable to a decomposition approach. For valid fixed values for variables y and x, the resulting subproblem decomposes into subsystems which are actually instances of shortest path problems, one for each $(i,j)\in W$. Hence a coordination scheme, akin Benders decomposition method (Benders, 1962) which iteratively proposes valid values for variables y and x and solves shortest path instance problems, has a great appeal, since it will most likely surpass an approach that directly solves the whole formulation (1)–(10) at once. It is also important to remark that this model is closely related to the model displayed in Camargo et al. ?, with several simplifications due to the size of the problems and instances under analysis.

3. Benders decomposition algorithms

The Benders decomposition technique (Benders, 1962) is a classical exact method suitable to solve mixed integer linear programs with a stair-case matrix structure. In general terms, the method partitions the original problem into two simpler problems of smaller dimensions: a master problem (MP) and a subproblem (SP). The MP is a relaxed version of the original problem having only its integer variables and their respective constraints, but having the continuous variables projected out. These variables are replaced by an auxiliary variable, responsible for sub-estimating the objective function of the projected subsystem, and by associated cutting planes known as Benders cuts. While the SP is the original problem with the integer variables temporarily fixed by the MP. The algorithm iterates by solving the MP followed by the SP, while Benders cuts are separated from the SP and added to the MP at each iteration, until the lower bound (LB) and the upper bound (UB) converge to an optimal solution, if one exists. The LB is provided by the MP, whereas a UB is readily available from the SPs' solutions.

The Benders decomposition technique has been successfully applied to different problems and other hub location variants (Geoffrion and Graves, 1974; Magnanti and Wong, 1981; Birge and Louveaux, 1988; Leung et al., 1990; Cordeau et al., 2000; Costa, 2005; Gelareh and Nickel, 2011; Contreras et al., 2011; Gelareh et al., 2015), being its performance closely related to the problem's structure, to how easily the SP is solved, and to the model's linear programming relaxation. A closer look at formulation (1)–(10) reveals that it shares these features. Hence, in this section, a Benders reformulation for the gateway hub location problem is shown, as well as a Benders algorithm to solve it. Further, a specialized efficient procedure that selects suitable dual optimal values among the multiple possible ones due to the dual SP's degeneracy is presented. This procedure is capable generates strong Benders cuts for the MP.

3.1. Benders subproblem and master problem

Let $\mathbb{Y} = \{(y,x)\in \mathbb{B}^{|H\cup G|\times|E|}\}$ be the set of feasible integer solutions associated to variables y and x for formulation (1)–(10). After parameterizing variables y and x, i.e. for $(\bar{y},\bar{x})\in\mathbb{Y}$, the following primal SP (PSP) is obtained:

$$\text{(PSP)}\; v(\bar{y},\bar{x}) = \min \sum_{(i,j)\in W}\sum_{(u,v)\in A} w_{ij}c_{uv}f_{uv}^{ij} \tag{11}$$

$$\text{s.t.}: \sum_{(i,v)\in A} f_{iv}^{ij} = 1 \qquad \forall\, (i,j)\in W \tag{12}$$

$$-\sum_{\substack{(v,u)\in A\\ j\neq v}} f_{vu}^{ij} + \sum_{\substack{(u,v)\in A\\ i\neq v}} f_{uv}^{ij} = 0 \qquad \forall\, (i,j)\in W, u\in N\colon i\neq u, j\neq u \tag{13}$$

$$-\sum_{(u,j)\in A} f_{uj}^{ij} = -1 \qquad \forall\, (i,j)\in W \tag{14}$$

$$-f_{uv}^{ij} - f_{vu}^{ij} \geq -\bar{x}_{uv} \qquad \forall\, (i,j)\in W, (u,v)\in E \tag{15}$$

$$-\sum_{(u,v)\in A} f_{uv}^{ij} \geq -\bar{y}_u \qquad \forall\, (i,j)\in W, u\in H\cup G \tag{16}$$

$$f_{uv}^{ij} \geq 0 \qquad \forall\, (i,j)\in W, (u,v)\in A \tag{17}$$

Let $\pi_{iju}\in\mathbb{R}$, and $\beta_{ijuv}\geq 0$ and $\rho_{iju}\geq 0$ be the dual variables associated with constraints (12)–(14), and (15) and (16) respectively. Then the dual SP (DSP) can be written as:

$$\text{(DSP)}\; v(\bar{y},\bar{x}) = \max \sum_{(i,j)\in W}\left[\pi_{iji} - \pi_{ijj} - \sum_{u\in H\cup G}\bar{y}_u\rho_{iju} - \sum_{(u,v)\in E}\bar{x}_{uv}\beta_{ijuv}\right] \tag{18}$$

$$\text{s.t.}: \pi_{iju} - \pi_{ijv} - \rho_{iju} \leq w_{ij}c_{uv} \qquad \forall\, (i,j)\in W, (u,v)\in A^{LH}\cup A^{HG} \tag{19}$$

$$\pi_{iju} - \pi_{ijv} - \beta_{ijuv} - \rho_{iju} \leq w_{ij}c_{uv} \qquad \forall\, (i,j)\in W, (u,v)\in A^{H}\cup A^{G}\colon u<v \tag{20}$$

$$\pi_{iju} - \pi_{ijv} - \beta_{ijvu} - \rho_{iju} \leq w_{ij}c_{uv} \qquad \forall\, (i,j)\in W, (u,v)\in A^{H}\cup A^{G}\colon u>v \tag{21}$$

$$\pi_{iju} - \pi_{ijv} \leq w_{ij}c_{uv} \qquad \forall\, (i,j)\in W, (u,v)\in A^{L} \tag{22}$$

Observe that the feasible space (19)–(22) is invariable for any $(\bar{y},\bar{x})$ value, and that the null vector is always a feasible solution to DSP, since $w_{ij}c_{uv}\geq 0$ for any $(i,j)\in W$ and $(u,v)\in A$. Further, the DSP is either bounded or unbounded, which, from strong duality, implies that either the PSP is feasible or infeasible, respectively. Hence, it is important to recognize when a $(\bar{y}, \bar{x}) \in \Upsilon$ vector renders into a feasible PSP, i.e. into a bounded DSP. The condition under which such vector exists is given by Proposition 1.

Proposition 1. *The primal and dual SPs are feasible and bounded, for any $(\bar{y}, \bar{x}) \in \Upsilon$ such that constraints (7), $\sum_{u\in G:R(u)=r} y_u \geq 1$ for all $r \in \mathscr{R}$ (i.e. there is at least one gateway per region), and $y_u \leq y_v$, for all $u \in G$ and $v \in H$, such that $\phi(u) = \phi(v)$ (i.e. a gateway can only be installed if its respective associated hub is also set), are respected.*

Proof. Since by the problem's definition about the pre-existence of the local arcs given by set A^l, there is at least one path for each local flow $(i, j) \in W^l$ within its region. Further constraints $\sum_{u\in G:R(u)=r} y_u \geq 1$ for all $r \in \mathscr{R}$, and (7) assure that all regions has at least an installed gateway, and are also linked forming a connected graph, i.e. there is a path consisted of gateway arcs connecting any pair of regions. Moreover as constraints $y_u \leq y_v$ for all $u \in G$ and $v \in H$, such that $\phi(u) = \phi(v)$, guarantee that a gateway can only exists if its respective associated hub is also installed, and, by the problem's definition about the local arc set A^l that establishes that there is always a local arc from a local airport to a installed hub, then there is at least a path for any global flow $(i, j) \in W^G$. As $w_{ij}c_{uv}$ are finite for $(i, j) \in W$ and $(u, v) \in A$, and due to constraints (12)-(16) then any feasible solution to the PSP must be bounded. Hence, by strong duality, the DSP must also be feasible and bounded.

Let $\mathbb{D}$ be the set of extreme points associated to DSP. It follows from Proposition 1 that the whole dual SP can then be expressed as:

$$v(\bar{y}, \bar{x}) = \max_{(\pi,\beta,\rho)\in\mathbb{D}} \sum_{(i,j)\in W} \left[\pi_{iji} - \pi_{ijj} - \sum_{u\in H\cup G} \bar{y}_u\rho_{iju} - \sum_{(u,v)\in E} \bar{x}_{uv}\beta_{ijuv}\right]$$

which, with the help of an auxiliary variable $\eta \geq 0$ to sub-estimate the routing costs, allows to reformulate formulation (1)-(10) as the following Benders MP:

$$(\text{BMP})\min \sum_{u\in H\cup G} a_u y_u + \sum_{(u,v)\in E} q_{uv}x_{uv} + \eta$$
$$\text{s.t.}: (7) - (9) \tag{23}$$

$$\eta \geq \sum_{(i,j)\in W} \left[\bar{\pi}_{iji} - \bar{\pi}_{ijj} - \sum_{u\in H\cup G} \bar{\rho}_{iju}y_u - \sum_{(u,v)\in E} \bar{\beta}_{ijuv}x_{uv}\right] \quad \forall\ (\bar{\pi}, \bar{\beta}, \bar{\rho}) \in \mathbb{D} \tag{24}$$

$$\sum_{\substack{u\in G\\ R(u)=r}} y_u \geq 1 \quad r \in \mathscr{R} \tag{25}$$

$$y_u \leq y_v \quad \forall\ u \in G, v \in H: \phi(u) = \phi(v) \tag{26}$$

$$\eta \geq 0 \tag{27}$$

Constraints (24) are known as Benders optimality cuts. There is one associated to each extreme point of the feasibility space of the DSP. As established by Proposition 1, the DSP is always bounded due to constraints (7) and (25) and (26), ergo there is no need for Benders feasibility cuts associated with the DSP's extreme rays to be added to the BMP. However if those constraints were to be disregarded, then Benders feasibility cuts like the following would be required:

$$0 \geq \sum_{(i,j)\in W} \left[\bar{\pi}_{iji} - \bar{\pi}_{ijj} - \sum_{u\in H\cup G} \bar{\rho}_{iju}y_u - \sum_{(u,v)\in E} \bar{\beta}_{ijuv}x_{uv}\right] \quad \forall\ (\bar{\pi}, \bar{\beta}, \bar{\rho}) \in \mathbb{E} \tag{28}$$

where $\mathbb{E}$ is the set of extreme rays associated with the DSP's feasibility space. In this study, both approaches are evaluated and assessed on the computational experiments. To aid in resolution, some auxiliary valid inequalities are also added to the BMP:

$$x_{uv} \leq y_u \quad \forall\ (u, v) \in E \tag{29}$$

$$x_{uv} \leq y_v \quad \forall\ (u, v) \in E \tag{30}$$

$$\sum_{(u,v)\in E} x_{uv} + \sum_{(v,u)\in E} x_{vu} \geq y_u \quad \forall\ u \in G \tag{31}$$

$$\sum_{\substack{(u,v)\in E\\ R(u)=r\wedge R(v)\neq r}} x_{uv} + \sum_{\substack{(v,u)\in E\\ R(v)\neq r\wedge R(u)=r}} x_{vu} \geq 1 \quad \forall\ r \in \mathscr{R} \tag{32}$$

$$\sum_{\substack{(u,v)\in E\\ R(u)\neq R(v)}} x_{u,v} \geq n_{\mathscr{R}} - 1 \tag{33}$$

Constraints (29) and (30) ensure that an inter-hub or inter-gateway connection is established only if the respective associated hubs or gateways are also set. Constraints (31) guarantee that each installed gateway is connected to at least another gateway. Constraints (32) insure that each region is connected to at least another region. Finally, constraint (33) determines that at least $(n_{\mathscr{R}} - 1)$ inter-gateway connections linking different regions are installed.

Problem (1)-(10) was then reformulated into an equivalent mixed integer program with fewer variables, having only the integer variables y and x, and one continuous variable η. Though the BMP has a smaller dimension than the original problem, it has now two constrain sets with an exponential size that must be managed in a suitable fashion. All but a few of the Benders optimality constraints and the SECs are initially disregarded to be iteratively added to the BMP, on demand, till an optimal solution is attained for the original problem.

3.2. A basic benders decomposition algorithm outline

Let UB and LB be the current upper and lower bounds, respectively, and h the current iteration index, $(\bar{y}^h, \bar{x}^h, \bar{\eta}^h)$ and $v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h)$ the current optimal solutions for the BMP and the DSP for iteration h, respectively. Let also $\bar{\mathbb{D}}$, $\bar{\mathbb{E}}$, and $\bar{\mathbb{S}}$ be the restricted sets of extreme points and rays of the DSP's feasible space, and of disconnected components within G generated so far, up to iteration h. An outline of the Benders decomposition technique steps is presented in Algorithm 1. This algorithm results in two different versions, called here as Alg-1-v1 and Alg-1-v2. While in Alg-1-v1 the DSP is solved at each iteration, in Alg-1-v2, a procedure detects whether the solution from $(\bar{y}, \bar{x})$ results in an infeasible or feasible solution to formulation (1)–(10) and the DSP is solved only when a feasible $(\bar{y}^h, \bar{x}^h)$ solution to formulation (1)–(10) is generated. Although both versions consider the addition of optimality Benders cuts, Alg-1-v1 adds feasibility Benders cut (28) and Alg-1-v2 adds SECs (7) when BMP solution is infeasible. The performance of Algorithm 1 is greatly affected by the computational effort spend on solving the BMP and DSP, and the total number of iterations required to attain optimality. These issues are addressed in the next sections.

Algorithm 1. Basic Benders decomposition

UB $\leftarrow +\infty$, LB $\leftarrow -\infty$, stop $\leftarrow$ **false**, $\bar{\mathbb{D}} \leftarrow \varnothing$, $\bar{\mathbb{E}} \leftarrow \varnothing$, $\bar{\mathbb{S}} \leftarrow \varnothing$
while (stop = **false**) **do**
  {solve MP}
  (LB, $\bar{y}^h, \bar{x}^h, \bar{\eta}^h$) $\leftarrow$ BMP($\bar{\mathbb{D}}, \bar{\mathbb{E}}$ or $\bar{\mathbb{S}}$)
  if (UB = LB) **then**
    stop $\leftarrow$ **true**
  else
    Alg-1-v1: {solve DSP}
    $v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h) \leftarrow$ DSP($\bar{y}^h, \bar{x}^h$)
    if ($v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h) < \infty$) **then**
      {bounded DSP}
      {add optimality Benders cuts}
      $\mathbb{D} \leftarrow \mathbb{D} \cup \{\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h\}$
      UB = min(UB, LB - $\bar{\eta}^h + v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h)$)
    else
      $\bar{\mathbb{E}} \leftarrow \bar{\mathbb{E}} \cup \{(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h)\}$
    end if
    OR
    Alg-1-v2: {test the feasibility of BMP solution}
    if BMP solution is infeasible **then**
      {add SECs}
      $s \leftarrow$ find disconnected components within $\{u \in G : \bar{y}_u^h = 1\}$
      $\bar{\mathbb{S}} \leftarrow \bar{\mathbb{S}} \cup \{s\}$
    else
      {solve DSP}
      $v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h) \leftarrow$ DSP($\bar{y}^h, \bar{x}^h$)
      {bounded DSP}
      {add optimality Benders cuts}
      $\mathbb{D} \leftarrow \mathbb{D} \cup \{\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h\}$
      UB = min(UB, LB - $\bar{\eta}^h + v(\bar{\pi}^h, \bar{\beta}^h, \bar{\rho}^h)$)
    end if
  end if
  $h \leftarrow h + 1$
end while

3.3. Adding Pareto Optimal cuts

One way to speed up the BMP's solution and shorten the required number of iterations for the Benders decomposition algorithm to reach optimality is by generating stronger, non dominated Benders cuts or Pareto-optimal cuts (Magnanti and Wong, 1981). A cut constructed from a dual solution $(\bar{\pi}^1, \bar{\beta}^1, \bar{\rho}^1)$ is said to dominate a cut assembled from another dual solution $(\bar{\pi}^2, \bar{\beta}^2, \bar{\rho}^2)$ if and only if $\sum_{(i,j)\in W}\left[\bar{\pi}_{iji}^1 - \bar{\pi}_{ijj}^1 - \sum_{u\in H\cup G}\bar{\rho}_{iju}^1 y_u - \sum_{(u,v)\in E}\bar{\beta}_{ijuv}^1 x_{uv}\right] \geq \sum_{(i,j)\in W}\left[\bar{\pi}_{iji}^2 - \bar{\pi}_{ijj}^2 - \sum_{u\in H\cup G}\bar{\rho}_{iju}^2 y_u - \sum_{(u,v)\in E}\bar{\beta}_{ijuv}^2 x_{uv}\right]$, for all $(y, x) \in \curlyvee$ with a strict inequality for at least one vector. A cut is said to be Pareto Optimal if it is not dominated by any other cut. To separate a Benders Pareto Optimal cut, Magnanti and Wong (1981) use a core-point or a reference point $(y^c, x^c) \in ri(\mathbb{P})$ belonging to the relative interior of the polyhedron $\mathbb{P} = \{(y, x)$: (7), (25), (26), $0 \leq y_u \leq 1, \forall\, u \in H \cap G$, and $0 \leq x_{uv} \leq 1, \forall\, (u, v) \in E\}$ when solving an additional dual SP, besides the regular DSP. In other words, two different SPs are solved at each iteration: on associated to the current BMP's solution $(\bar{y}^h, \bar{x}^h)$, and another related to the core point (y^c, x^c).

As this additional dual SP has a further dense equality constraint, it poses as a much harder problem to solve than the DSP and prone to numerical instabilities. Papadakos (2008) then proposes a lighter version to generate Benders Pareto Optimal cuts by solving the same DSP but with the core point $(y^c, x^c) \in ri(\mathbb{P})$ in place of the current BMP's solution $(\bar{y}^h, \bar{x}^h)$ instead, or:

$$(\text{PDSP}) \max_{(\pi,\beta,\rho)\in\mathbb{D}} \sum_{(i,j)\in W}\left[\pi_{iji} - \pi_{ijj} - \sum_{u\in H\cup G} y_u^c \rho_{iju} - \sum_{(u,v)\in E} x_{uv}^c \beta_{ijuv}\right] \tag{34}$$

but updating the core point at each iteration in which the DSP renders a bounded solution. The core point is updated by a linear convex combination of the current BMP's solution $(\bar{y}^h, \bar{x}^h)$ with the core point or $(y^c, x^c) = \lambda(y^c, x^c) + (1-\lambda)(\bar{y}^h, \bar{x}^h)$, in which $0 < \lambda < 1$. Empirically $\lambda = 0.5$ provides the best overall results (Papadakos, 2008; Mercier et al., 2005). An initial core point is then required for starting the Benders decomposition algorithm. A valid one which respects the definition of $ri(\mathbb{P})$ can be given as:

$$(\text{ICP})\, y_u^c = \frac{1}{|\{u \in G: R(u) = r\}|} \qquad \forall\, r \in \mathscr{R} \tag{35}$$

$$y_u^c = y_v^c \qquad \forall\, u \in H,\, v \in G: \phi(u) = \phi(v) \tag{36}$$

$$x_{uv}^c = \min(y_u^c, y_v^c) \qquad \forall\, (u, v) \in E \tag{37}$$

Proposition 2. *The vector (y^c, x^c) given by ICP is a valid core point, i.e. $(y^c, x^c) \in ri(\mathbb{P})$.*

Proof. *By construction, constraints (25) and (26) are respected by equalities (35) and (36), respectively, and $0 < (y^c, x^c) < 1$. Finally, the SECs are also attended since $\sum_{(u,v)\in E^G u\in S, v\in G\setminus S} \min(y_u^c, y_v^c) + \sum_{(v,u)\in E^G u\in G\setminus S, v\in S} \min(y_u^c, y_v^c) \geq y_k^c + y_m^c - 1$ for all $k \in G\setminus S$, $m \in S$ such that $S \subset \mathbb{S}$ and $R(k) \neq R(m)$. Recall the k and m belong to different regions ($R(k) \neq R(m)$), hence their associated edge connection, i.e variable x_{km}^c if $k < m$, or x_{mk}^k if $k > m$, will appear on left hand side of its associated SEC. Further as $0 < y_k^c < 1$ and $0 < y_m^c < 1$ and $x_{km}^c = \min(y_k^c, y_m^c)$ by construction, then $\min(y_k^c, y_m^c) > y_k^c + y_m^c - 1$ for any pair of gateways located in different regions.*

Algorithm 2 shows an outline of the Papadakos' Benders decomposition approach. Note that the PDSP is solved before the BMP, while using the current core point, and that, whenever the DSP is bounded, the core point is updated by a linear convex combination. Furthermore, this algorithm expound two variants of the Benders decomposition method, Alg-2-v1 and Alg-

2-*v*2. *Similarly to Algorithm 1, Alg-2-v1 solves the DSP at each iteration, adding feasibility Benders cuts when necessary; while Alg-2-v2 solves the DSP only when an infeasible* $(\bar{y}^h, \bar{x}^h)$ *solution to formulation* (1)-(10) *is generated, separating SECs to be added to BMP. This particular solution ordering might speed up the convergence of the Benders decomposition algorithm* (Papadakos, 2008).

Algorithm 2. Papadakos based Benders decomposition

```
UB ← +∞, LB ← −∞, stop ← false, D̄ ← ∅, Ē ← ∅, S̄ ← ∅
bounded ← true
(y^c, x^c) ← ICP
while (stop = false) do
  if (bounded = true) then
    {solve PDSP}
    v(π̄^h, β̄^h, ρ̄^h) ← PDSP(y^c, x^c)
    D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
  end if
  {solve MP}
  (LB, ȳ^h, x̄^h, η̄^h) ← BMP(D̄, Ē or S̄)
  if (UB = LB) then
    stop ← true
  else
    Alg-2-v1: {solve DSP}
    v(π̄^h, β̄^h, ρ̄^h) ← DSP(ȳ^h, x̄^h)
    if (v(π̄^h, β̄^h, ρ̄^h) < ∞) then
      {bounded DSP}
      bounded ← true
      (y^c, x^c) ← λ(y^c, x^c) + (1 − λ)(ȳ^h, x̄^h),
      {add optimality Benders cuts}
      D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
      UB = min(UB, LB - η̄^h + v(π̄^h, β̄^h, ρ̄^h))
    else
      {unbounded DSP}
      bounded ← false
      Ē ← Ē ∪ {(π̄^h, β̄^h, ρ̄^h)}
    end if
```

Continuation of algorithm 2 Papadakos based Benders decomposition

```
    OR
    Alg-2-v2: {test the feasibility of BMP solution}
    if BMP solution is infeasible then
      {add SECs}
      s ← find disconnected components within {u ∈ G : ȳ^h_u = 1}
      S̄ ← S̄ ∪ {s}
    else
      {solve DSP}
      v(π̄^h, β̄^h, ρ̄^h) ← DSP(ȳ^h, x̄^h)
      {bounded DSP}
      bounded ← true
      (y^c, x^c) ← λ(y^c, x^c) + (1 − λ)(ȳ^h, x̄^h),
      {add optimality Benders cuts}
      D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
      UB = min(UB, LB - η̄^h + v(π̄^h, β̄^h, ρ̄^h))
    end if
  end if
  h ← h + 1
end while
```

3.4. Adding multiple benders cuts

For feasible integer solutions $(x, y) \in \Upsilon$ for formulation (1)–(10), the PSP can be decomposed into $|W|$ independent subproblems, one for each $(i, j) \in W$. Each independent subproblem consists of a minimum shortest problem which can be easily solved by Dijkstra's algorithm. Further it allows the separation of Benders optimality cuts from each independent subproblem that can then be added to the BMP. This greatly reduces the number of iterations required by the method to reach optimality (Birge and Louveaux, 1988). Let $\mathbb{D}_{ij}$ be the set of extreme points associated to DSP regarding the independent subproblem (i, j), which, after redefining the variables η_{ij} accordingly, allows the BMP to be rewritten into a stronger form:

$$\min \sum_{u\in H\cup G} a_u y_u + \sum_{(u,v)\in E} q_{uv} x_{uv} + \sum_{(i,j)\in W} \eta_{ij}$$

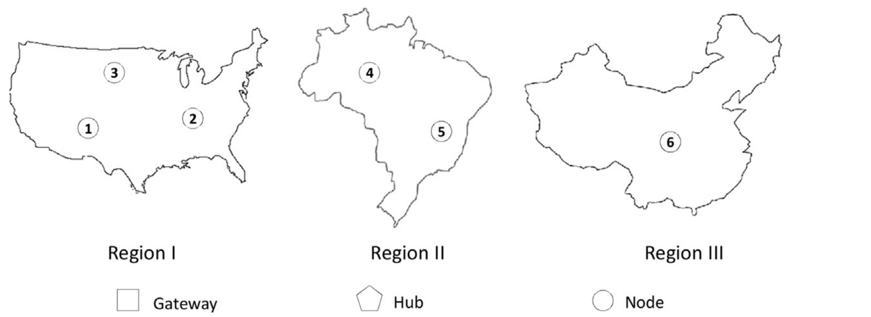
$$\text{s.t.}: (7) - (9) \text{ and } (25) \text{ and } (26) \tag{38}$$

$$\eta_{ij} \geq \bar{\pi}_{iji} - \bar{\pi}_{ijj} - \sum_{u\in H\cup G} \bar{\rho}_{iju} y_u - \sum_{(u,v)\in E} \bar{\beta}_{ijuv} x_{uv} \qquad \forall\ (i, j) \in W, (\bar{\pi}, \bar{\beta}, \bar{\rho}) \in \mathbb{D}_{ij} \tag{39}$$

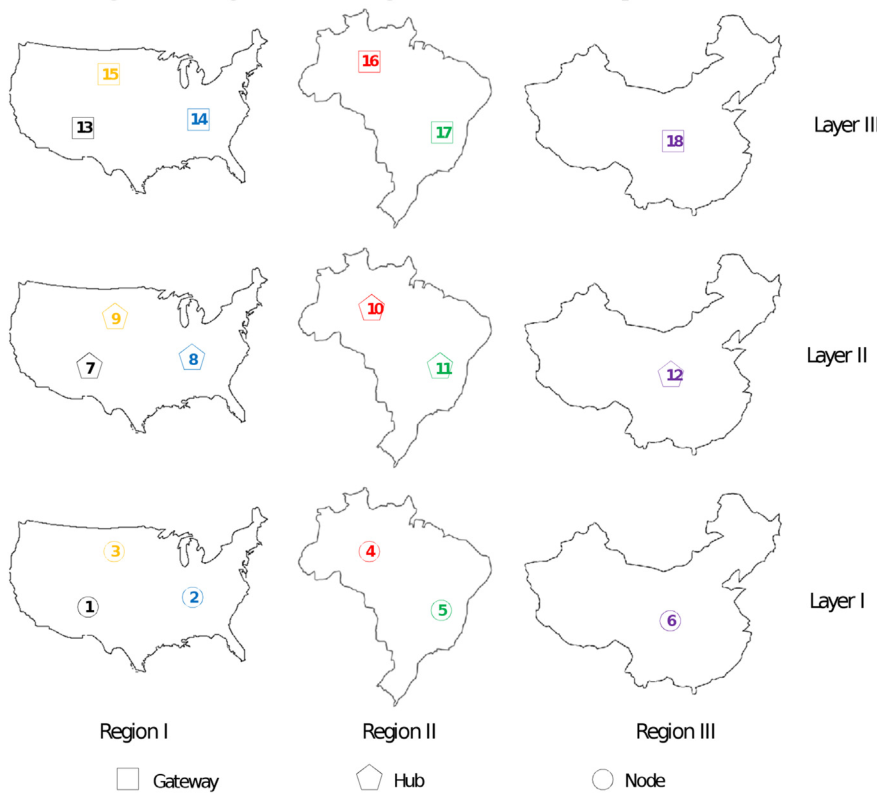
$$\eta_{ij} \geq 0 \qquad \forall\ (i, j) \in W \tag{40}$$

3.5. Solving benders subproblem

Regarding the subproblem to find SECS within the installed hubs and gateways at an iteration h, a maximum flow problem (Ahuja et al., 1993, page 240), or the Tarjan's depth-first search (Tarjan, 1972) to detect strong components can be used. Further, instead of relying on a linear programing solver to solve the DSP or the $|W|$ independent subproblems – one associated to each $\mathbb{D}_{ij}$ – induced by feasible solutions $(y, x) \in \Upsilon$ to formulation (1)–(10), a specialized procedure is devised to exploit the shortest path structure of the primal subproblem to



(a) Original setting with three regions and six local airports.



(b) Setting with three layers.

Fig. 1. Example with the original setting and with a three layer configuration.

efficiently generate dual optimal values. Let $\mathscr{H}_u^{ij}$ be the length of the shortest path from i to node u on the shortest path from i to j for $(i, j) \in W$ induced by a feasible integer solution $(\bar{y}^h, \bar{x}^h) \in \Upsilon$ to formulation (1)–(10) at iteration h. Note that $\mathscr{H}_j^{ij} = \sum_{(u,v) \in A} w_{ij} c_{uv} f_{uv}^{ij}$ corresponds to the shortest path from i and j for $(i, j) \in W$ in the rendered network by $(\bar{y}^h, \bar{x}^h)$. Let also $\mathscr{T}_u^{ij}$ be the length of the shortest path from u to node j in the shortest path from i to j, but considering all hubs and gateways, and inter-hub and inter-gateway connections installed, i.e. all ones vector $(y, x) = \mathbf{1}$. Define $\delta_u^{ij} = \mathscr{H}_j^{ij} - \mathscr{T}_u^{ij}$, for all $u \in N$. Given these definitions, the optimal dual solution for $(i, j) \in W$ can be calculated as:

$$\pi_{iji} = 0 \tag{41}$$

$$\pi_{ijj} = \mathscr{H}_j^{ij} \tag{42}$$

$$\pi_{iju} = \min(\mathscr{H}_u^{ij}, \delta_u^{ij}) \qquad \forall\ u \in N: u \neq i \wedge u \neq j \tag{43}$$

$$\rho_{iju} = \max(0, \pi_{iju} - \pi_{ijv} - w_{ij} c_{uv}) \qquad \forall\ (u, v) \in A: u \in H \cap G \tag{44}$$

$$\beta_{ijuv} = \max(0, \pi_{iju} - \pi_{ijv} - \rho_{iju} - w_{ij} c_{uv}) \qquad \forall\ (u, v) \in A^H \cap A^G: u < v \tag{45}$$

$$\beta_{ijvu} = \max(0, \pi_{iju} - \pi_{ijv} - \rho_{iju} - w_{ij} c_{uv}) \qquad \forall\ (u, v) \in A^H \cap A^G: u > v \tag{46}$$

Proposition 3. Equations 41–46 *compute optimal dual values for* $\mathbb{D}_{ij}$ *for feasible integer solutions* $(y, x) = (\bar{y}^h, \bar{x}^h)$ *to formulation* (1)-(10) *at iteration h.*

Proof. *By construction*, equations (44)–(46) *insure that the set* (π, ρ, β) *is feasible for* $\mathbb{D}_{ij}$. *Since* equations (41) and (42) *set* $\pi_{iji} = 0$ *and* $\pi_{ijj} = \mathscr{H}_j^{ij}$ *by definition, respectively, this allows to interpret variables* β_{ijuv} *and* ρ_{iju} *as the reduction in the shortest path length from node i to j if the inter-hub or inter-gateway connection* $(u, v) \in E$, *or the hub or gateway airport* $u \in H \cap G$

are installed. Hence, when $\bar{x}_{uv}^h = 1$ *or* $\bar{y}_u^h = 1$ *imply that* $\beta_{ijuv} = 0$ *or* $\rho_{iju} = 0$ *because of the complementary slackness condition to ensure that* $\pi_{iji} = 0$ *and* $\pi_{ijj} = \mathscr{H}_j^{ij}$*, guarantee therefore that the optimal dual solution for* $\mathbb{D}_{ij}$ *will be* $\mathscr{H}_j^{ij}$.

3.6. Repairing infeasible primal solutions

Whenever the DSP is unbounded at an iteration h, no optimality Benders cuts are separated, no core point is updated, no Pareto Optimal cuts are generated, and no improvement on the Benders decomposition algorithm's lower bound might be perceived on the next iteration. Hence no contribution to the method's convergence is observed. To dwindle this effect, a simple, but effective repair procedure to find a feasible solution from a $(\bar{y}, \bar{x})$ infeasible solution to formulation (1)–(10) is developed. For iterations with infeasible network generated by the BMP, beyond the separation of SECs, the repair procedure is called followed by the specialized algorithm to solve the DSP so that optimality Benders cuts can be separated. With this procedure, the core points can be updated at each iteration and consequently, Pareto Optimal Benders cuts can be added even when a $(\bar{y}, \bar{x})$ infeasible solution to formulation (1)–(10) is found.

Since all local airports are directly connected, the origin of an infeasible solution is from global demand. Therefore, make a network feasible implies in guarantee that all regions are connected. Since the regions are connected by gateways arcs, this subproblem considers the arcs $(u, v) \in A^G$. The global demand, $(i, j) \in W^g$, is aggregated on gateways, resulting in an auxiliary demand matrix, W^{aux}. Each node $i \in L$ is associated to the nearest gateway $u \in G$, such that $R(i) = R(u)$, and the demand of this node $i \in I$ is added on the gateway $u \in G$ demand. To illustrate, assume that there is a demand of 10 units from node 1 ($R(1) = 1$ and closer to gateway 14, $R(14) = 1$) to node 2 ($R(2) = 2$ and closer to gateway 17, $R(17) = 2$). Thus, the new demand from gateway 14 to gateway 17 will be at least equal 10 units.

Using the subdigraph $\mathbb{G}' = (G, A^G)$ of Section 2, and redefining the unitary transportation cost c_{uv} for an infeasible $(\bar{y}^h, \bar{x}^h)$ solution to formulation (1)–(10) at iteration h as:

$$c_{uv} = \begin{cases} q_{uv} & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{uv}^h = 0 \wedge \overline{y_u}^h = 1 \wedge \overline{y_v}^h = 1 \\ q_{uv} + a_u & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{uv}^h = 0 \wedge \overline{y_u}^h = 0 \wedge \overline{y_v}^h = 1 \\ q_{uv} + a_v & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{uv}^h = 0 \wedge \overline{y_u}^h = 1 \wedge \overline{y_v}^h = 0 \\ q_{uv} + a_u + a_v & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{uv}^h = 0 \wedge \overline{y_u}^h = 0 \wedge \overline{y_v}^h = 0 \\ 0 & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{uv}^h = 1 \\ 0 & \forall\ (u, v) \in A^G: ((u, v) \vee (v, u)) \in E \wedge \bar{x}_{vu}^h = 1 \end{cases}$$

Then, for each $(i, j) \in W^{aux}$, a Dijkstra's algorithm is called on the a subdigraph $\mathbb{G}' = (G, A^G)$. The $(\bar{y}^h, \bar{x}^h)$ solution is repaired at the end, by activating the hubs and gateways, and inter-hub and inter-gateway connections that are part of the attained shortest path, but were not originally present in the $(\bar{y}^h, \bar{x}^h)$ solution. The specialized algorithm of Section 3.5 calculates the optimal dual values for the repaired solution and a new optimality Benders cut is added to the BMP, as can be seen in Algorithm 3. Note that if Pareto Optimal Benders cuts are not considered, we can have two versions of Algorithm 3, Alg-3-v1 and Alg3-v2. In Alg-3-v1, only optimality Benders cuts and SECs are separated, while in Alg-3-v2, Pareto Optimal Benders cuts are also added at each iteration.

Algorithm 3. Repair Benders decomposition

```
UB ← +∞, LB ← −∞, stop ← false, D̄ ← ∅, Ē ← ∅, S̄ ← ∅
(y^c, x^c) ← ICP
while (stop = false) do
   {solve PDSP}
   v(π̄^h, β̄^h, ρ̄^h) ← PDSP(y^c, x^c)
   D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
   {solve MP}
   (LB, ȳ^h, x̄^h, η̄^h) ← BMP(D̄, Ē or S̄)
   if (UB = LB) then
      stop ← true
   else
      {test the feasibility of BMP solution}
      if BMP solution is infeasible then
         {add SECs}
         s ← find disconnected components within {u ∈ G : ȳ_u^h = 1}
         S̄ ← S̄ ∪ {s}
         {repair solution}
         (ȳ^h, x̄^h) ← Repair(ȳ^h, x̄^h)
         {solve DSP}
         v(π̄^h, β̄^h, ρ̄^h) ← DSP(ȳ^h, x̄^h)
         D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
      else
         {solve DSP}
         v(π̄^h, β̄^h, ρ̄^h) ← DSP(ȳ^h, x̄^h)
         {bounded DSP}
         {add optimality Benders cuts}
         D ← D ∪ {π̄^h, β̄^h, ρ̄^h}
      end if
      (y^c, x^c) ← λ(y^c, x^c) + (1 − λ)(ȳ^h, x̄^h),
      UB = min(UB, LB - η̄^h + v(π̄^h, β̄^h, ρ̄^h))
   end if
   h ← h + 1
end while
```

Table 1
Description of the global instances created.

Instances	# Nodes	# Hubs	# Gateways	# Regions
global-12	12	11	11	8
global-19	19	17	17	10
global-29	29	24	24	14
global-37	37	28	28	15
global-43	43	32	32	15
global-48	48	35	35	15
global-59	59	40	40	15
global-74	74	48	48	16
global-100	100	60	60	18
global-141	141	67	67	19

4. Computational experiments

In order to test the proposed variants of the Benders decomposition for the GHLP, we have generated an instance set, called here as global set. To compose the instance set, 141 big cities have been selected. Each city has its latitude, longitude, population and Gross Domestic Product (GDP). Depending on its location, they have been divided in regions.

Assuming that p_i is the population of city i divided by 100000, g_i the factor representing the GDP of city i, d_{ij} the distance between cities i and j, and the demand from city i to city j w_{ij} is expressed as follows:

$$w_{ij} = p_i p_j g_i g_j e^{-0.01 d_{ij}} \tag{47}$$

Cities are placed in descending order of population, and then, the n most populous cities are selected to compose a new instance. Hence, if $k > m$, an instance of k nodes has all m nodes presented in the m nodes instance and also more $k - m$ different nodes. In other words, global-19 instance has the same nodes of global-12 instance and other 7 different cities. Different problems consisting of the first n nodes have been generated for n = 12, 19, 29, 37, 43, 48, 59, 74, 100 and 141. Table 1 relates the instance's name with its number of nodes, candidates to become a hub, candidates to become a gateway and regions.

Table 2
Five settings.

Values	Setting I	Setting II	Setting III	Setting IV	Setting V
Hub fixed cost	10^3	10^4	10^4	10^5	10^5
Gateway fixed cost	10^4	10^4	10^5	10^5	10^6
Weight of domestic arcs	10^{-1}	10^0	10^0	10	10
Weight of international arcs	10^3	10^3	10^4	10^4	10^5

On one hand, the fixed costs for an airport becoming hub or gateway do not differ from one airport to another. On the other hand, the fixed costs to install arcs between domestic or international airports are a weighted funciton of the arc lenght. The round of experiments aims to evaluate how instances become harder by adopting more aggressive fixed costs. Table 2 shows five different settings evaluated. In setting I, the lowest fixed costs are considered, while, in setting V, the largest fixed costs are taken into account. In settings II and IV, the cost of activating hubs and gateways is assumed to be the same. In settings I, III and V, activating gateways is more expensive than opening hubs. We always assume that installing hub arcs is cheaper than activating gateway arcs.

All computational tests have been carried out on a Dell PowerEdge T620 workstation, equipped with two Intel Xeon E5-2600v2 processors and 96 GB of RAM memory. Also, all the algorithms have been implemented in C+ + using Concert Technology (CPLEX 12.5).

We have generated 8 variants of the Benders decomposition by changing the methods used to solve the DSP. Alg-1-v2 and Alg-2-v2, in fact, become Alg-1-v2-C, Alg-1-v2-I, Alg-2-v2-C and Alg-2-v2-I. The letter "C" indicates that the DSP is solved by CPLEX, while the letter "I" points out the implementation of the specialized algorithm of Section 3.5. The monolithic version corresponds to equations 1–10.

Table 3
Results for smaller instances - Setting I.

Versions	global-12			global-19			global-29			global-37		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	0	–	0.00	3	–	0.00	26	–	0.00	349
Alg-1-v1	11	0.00	4	12	0.00	28	16	0.00	207	17	0.00	770
Alg-1-v2-C	13	0.00	3	17	0.00	23	23	0.00	189	27	0.00	1230
Alg-1-v2-I	12	0.00	0	16	0.00	5	22	0.00	31	18	0.00	45
Alg-3-v1	7	0.00	0	11	0.00	4	10	0.00	14	10	0.00	26
Alg-2-v1	8	0.00	3	9	0.00	18	11	0.00	161	10	0.00	494
Alg-2-v2-C	13	0.00	4	12	0.00	27	9	0.00	122	11	0.00	322
Alg-2-v2-I	13	0.00	2	13	0.00	14	12	0.00	67	14	0.00	169
Alg-3-v2	8	0.00	3	8	0.00	15	8	0.00	77	8	0.00	189

Table 4
Results for smaller instances - Setting II.

Versions	global-12			global-19			global-29			global-37		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	0	–	0.00	2	–	0.00	24	–	0.00	396
Alg-1-v1	9	0.00	3	13	0.00	29	15	0.00	193	17	0.00	1055
Alg-1-v2-C	14	0.00	3	20	0.00	31	24	0.00	213	32	0.00	984
Alg-1-v2-I	14	0.00	1	19	0.00	10	23	0.00	59	23	0.00	114
Alg-3-v1	10	0.00	1	15	0.00	11	13	0.00	35	13	0.00	68
Alg-2-v1	7	0.00	3	8	0.00	26	12	0.00	221	11	0.00	527
Alg-2-v2-C	13	0.00	4	12	0.00	27	10	0.00	145	13	0.00	326
Alg-2-v2-I	13	0.00	2	13	0.00	14	12	0.00	68	14	0.00	170
Alg-3-v2	8	0.00	3	8	0.00	15	8	0.00	77	8	0.00	191

Table 5
Results for smaller instances - Setting III.

Versions	global-12			global-19			global-29			global-37		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	0	–	0.00	13	–	0.00	241	–	0.00	959
Alg-1-v1	10	0.00	4	16	0.00	67	28	0.00	6513	35	0.09	–
Alg-1-v2-C	13	0.00	4	24	0.00	67	34	0.00	4228	40	0.08	–
Alg-1-v2-I	12	0.00	1	25	0.00	42	34	0.00	1179	39	0.00	7008
Alg-3-v1	8	0.00	1	41	0.00	19	24	0.00	1040	24	0.00	6218
Alg-2-v1	7	0.00	4	11	0.00	41	20	0.00	888	22	0.00	5513
Alg-2-v2-C	9	0.00	3	14	0.00	33	18	0.00	257	22	0.00	576
Alg-2-v2-I	10	0.00	2	22	0.00	29	24	0.00	144	23	0.00	317
Alg-3-v2	6	0.00	2	10	0.00	26	13	0.00	161	15	0.00	452

Table 6
Results for smaller instances - Setting IV.

Versions	global-12			global-19			global-29			global-37		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	0	–	0.00	14	–	0.00	275	–	0.00	1182
Alg-1-v1	10	0.00	5	19	0.00	262	34	0.00	93707	19	7.43	–
Alg-1-v2-C	14	0.00	5	29	0.00	390	46	0.00	72244	32	6.30	–
Alg-1-v2-I	26	0.00	10	66	0.00	1918	81	0.33	–	44	7.30	–
Alg-3-v1	16	0.00	8	50	0.00	1682	66	1.08	–	32	5.76	–
Alg-2-v1	8	0.00	4	15	0.00	74	19	0.00	1098	19	0.00	5287
Alg-2-v2-C	8	0.00	4	12	0.00	33	15	0.00	232	26	0.00	898
Alg-2-v2-I	8	0.00	2	16	0.00	20	26	0.00	158	33	0.00	439
Alg-3-v2	6	0.00	2	13	0.00	27	14	0.00	188	15	0.00	492

Table 7
Results for smaller instances - Setting V.

Versions	global-12			global-19			global-29			global-37		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	2	–	0.00	195	–	0.00	4647	–	0.00	20405
Alg-1-v1	9	0.00	4	18	0.00	116	28	0.00	8232	41	0.00	194300
Alg-1-v2-C	13	0.00	4	23	0.00	127	35	0.00	3493	51	0.00	231401
Alg-1-v2-I	16	0.00	3	42	0.00	365	70	0.00	41388	66	0.02	–
Alg-3-v1	12	0.00	3	39	0.00	362	61	0.00	26466	48	0.00	134488
Alg-2-v1	7	0.00	4	9	0.00	44	18	0.00	645	21	0.00	3436
Alg-2-v2-C	8	0.00	3	11	0.00	28	16	0.00	200	16	0.00	506
Alg-2-v2-I	10	0.00	2	12	0.00	18	18	0.00	107	19	0.00	290
Alg-3-v2	9	0.00	3	10	0.00	26	11	0.00	149	13	0.00	441

To assess the implemented versions, in Tables 3–12, we report the number of iterations required until convergence for the Benders proposed variants algorithms (# Iters), the total time required to attain an optimal solution (Time [s]), up to a limit of 259200 s (3 days), and the GAP when an optimal solution is not found within this time (GAP) were recorded. The symbol '-' in GAP column represents that no integer solution was found in the maximum time determined, and in time column that the algorithm ran until the maximum time allowed.

Two sets of experiments were realized. On the first one, comparison between all the variants of the Benders decomposition method is made for the smaller instances (until 37 airports). Tables (3)–(7) illustrate the results obtained for settings I, II, III, IV and V, respectively.

As can be seen on Tables (3)–(7), in all settings, except in setting V, the Monolithic version presents the best behavior for smaller instances (global-12 and global-19). For instances global-29 and global-37, there are always a variant of the Benders decomposition method that performs better than the Monolithic version. As fixed costs increase, so does the performance difference between the best version of Benders and the Monolithic version for a given instance. For example, while monolithic version requires 13 times more computational effort than Alg-3-v1 to solve global-37 instance in setting I, in setting V, monolithic version requires 70 times more computational effort than Alg-2-v2-I to solve the same instance.

It is possible to see that the instances become harder as fixed costs increase. The Monolithic version spends about 58 times longer to find an optimal solution for the global-37 instance in scenario V than in scenario I. Moreover, while the global-37 instance is solved within 26 s by Alg-3-v1 in setting I, in setting V, Alg-2-v2-I requires 290 s to converge to an optimal solution. When fixed costs are higher, the influence of transportation costs decreases and the problem becomes more combinatory.

Furthermore, the versions that consider the solution of DSP by using CPLEX and/or the addition of feasibility cuts instead of SECs (Alg-1-v1, Alg-1-v2-C, Alg-2-v1, Alg-2-v2-C) perform worse than the other versions. This implies that CPLEX requires more time to solve the DSP than the method described in 3.5. Although the versions Alg-2-v2-I and Alg-3-v1 are not able to find an optimal solution within the time limit for some instances in settings IV and V, they performed very well in settings I and II, in which fixed costs are lower. As expected, the separation of Pareto Optimal cuts presents a better behavior for harder instances.

As the best versions in setting I and II are Alg-1-v2-I and Alg-3-v1 and in settings III, IV and V, are Alg-2-v2-I and Alg-3-v2, the second

Table 8
Results for larger instances - Setting I.

Versions	global-43			global-48			global-59			global-74			global-100			global-141		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	1009	–	0.00	1879	–	0.00	3801	–	0.00	9882	–	0.00	45298	–	–	–
Alg-1-v2-I	24	0.00	166	44	0.00	941	37	0.00	3815	55	0.02	27149	21	–	–	37	6.42	–
Alg-3-v1	11	0.00	74	13	0.00	204	16	0.00	889	14	0.00	2523	13	0.07	–	8	1.48	–
Alg-2-v2-I	14	0.00	395	13	0.00	640	23	0.00	1590	56	0.00	23935	27	0.00	23031	34	0.00	59391
Alg-3-v2	8	0.00	388	9	0.00	709	8	0.00	1472	9	0.00	4545	8	0.00	14124	9	0.00	55291

Table 9
Results for larger instances - Setting II.

Versions	global-43			global-48			global-59			global-74			global-100			global-141		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	1230	–	0.00	2137	–	0.00	4463	–	0.00	10023	–	0.00	39442	–	–	–
Alg-1-v2-I	26	0.00	325	46	0.00	1753	42	0.00	9196	53	0.01	–	22	–	–	33	12.49	–
Alg-3-v1	14	0.00	280	17	0.00	567	22	0.00	7062	19	0.00	162188	11	3.54	–	6	8.06	–
Alg-2-v2-I	13	0.00	395	14	0.00	732	25	0.00	1814	30	0.00	5272	35	0.00	35198	32	0.00	65951
Alg-3-v2	8	0.00	391	9	0.00	725	9	0.00	1686	9	0.00	4418	10	0.00	17995	10	0.00	62404

Table 10
Results for larger instances - Setting III.

Versions	global-43			global-48			global-59			global-74			global-100			global-141		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	1859	–	0.00	3928	–	0.00	10301	–	0.00	37840	–	–	–	–	–	–
Alg-1-v2-I	46	0.00	63480	62	0.00	111016	46	2.97	–	62	4.63	–	22	–	–	26	90.60	–
Alg-3-v1	28	0.00	43694	28	0.00	70738	22	1.10	–	15	3.15	–	9	13.44	–	6	20.43	–
Alg-2-v2-I	31	0.00	631	27	0.00	1019	29	0.00	2896	36	0.00	1489	44	0.00	200848	32	0.27	–
Alg-3-v2	12	0.00	668	11	0.00	1178	16	0.00	3540	13	0.00	9154	14	0.00	83255	10	1.64	–

Table 11
Results for larger instances - Setting IV.

Versions	global-43			global-48			global-59			global-74			global-100			global-141		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	1968	–	0.00	4386	–	0.00	13407	–	0.00	39828	–	–	–	–	–	–
Alg-1-v2-I	44	17.95	–	54	18.64	–	41	25.68	–	53	28.39	–	24	–	–	29	88.01	–
Alg-3-v1	23	13.71	–	18	17.81	–	14	24.36	–	9	31.04	–	6	41.90	–	5	39.75	–
Alg-2-v2-I	23	0.00	640	32	0.00	1410	20	0.00	2669	40	0.00	24736	39	0.00	114895	26	2.55	–
Alg-3-v2	13	0.00	755	14	0.00	1520	12	0.00	2887	16	0.00	10778	15	0.00	56425	6	2.90	–

Table 12
Results for larger instances - Setting V.

Versions	global-43			global-48			global-59			global-74			global-100			global-141		
	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]	# Iters	GAP (%)	Time [s]
Monolithic	–	0.00	45241	–	0.00	52210	–	0.00	152610	–	0.01	–	–	–	–	–	–	–
Alg-1-v2-I	51	2.59	–	59	5.48	–	49	15.98	–	44	14.53	–	22	–	–	29	66.40	–
Alg-3-v1	34	2.38	–	28	4.47	–	17	9.25	–	10	15.75	–	7	28.00	–	4	40.96	–
Alg-2-v2-I	21	0.00	712	28	0.00	1940	32	0.00	11554	42	0.00	35208	28	6.37	–	29	9.97	–
Alg-3-v2	13	0.00	923	15	0.00	2128	16	0.00	10078	16	0.00	19112	15	0.00	126468	7	7.30	–

experiment compares these versions with the Monolithic version. As can be seen on Tables 8–12, the larger the instance, the more the performance of Alg-2-v2-I and Alg-3-v2 versions stand out. In setting I, which has the lowest fixed costs, Alg-3-v1 converges faster until global-74 instance. Alg-3-v2 outperforms the Alg-3-v1 on the global-59 instance in setting II. In the other settings, either Alg-2-v2-I or Alg-3-v2 present the best behavior. Again, it shows up that Pareto Optimal Benders cuts are effective on harder instances.

The fact that instances on setting V are harder than those on setting I, once more, stands out. For global-141 instance, while an optimal solution is determined in 55291 and in 62404 s in settings I and II, respectively, none of the versions can find an optimal solution within the maximum computational time allowed in the other settings.

Comparing versions Alg-2-v2-I and Alg-3-v2, we can observe that Alg-2-v2-I always requires more iterations to converge to an optimal solution. In setting III, for example, for global-100 instance, Alg-2-v2-I takes 3 times more iterations and almost 2.5 times more computational time to converge. However, not always Alg-3-v2 requires less time to find an optimal solution. For global-74 instance, in setting III, Alg-3-v2 takes 6 times more computational time to converge. This behavior can be explained by observing that in Alg-3-v2, Pareto Optimal Benders cuts are separated at each iteration, even though an infeasible $(\bar{y}^h, \bar{x}^h)$ solution to formulation (1)–(10) is generated. As these cuts are stronger, the number of iterations required to the convergence of the method decreases, but the time required in each iteration increases (PM becomes harder to solve and also we have to consider that the PDSP is solved at each iteration). Therefore, there is a trade-off between the time spent in each iteration and the necessary number of iterations to the convergence of the algorithm.

It is remarkable the better performance of Alg-2-v2-I and Alg-3-v2 compared to Monolithic version. For global-100 instance, for example, while Alg-2-v2-I and Alg-3-v3 converges to an optimal solution using less than 50% of the maximum time set in five settings, the monolithic version is able to find an optimal solution only for the first two settings (I and II).

5. Insights from this study

The problem proposed in this paper may act as a tool for providing insights into the current air network. First, empirical results for intermediate instances are presented for a variety of cost parameter values. Next, in order to exemplify some practical benefits of the work developed here, the biggest instance (global-141 instance) is chosen to be analyzed.

5.1. Measuring the influence of economies of scale and transportation costs on the optimal topologies

To illustrate the influence of the transportation costs in an optimal network, our formulation is solved to optimality by using different values for scale economies (α^H, α^G) and unitary operational costs (b^H, b^G). Our first analyses employ global-37, global-43 and global-48 instances in a series of numerical examples. Whereas it is not possible to make sweeping generalizations by deploying only three test problems, it is possible though to compare and contrast the features and characteristics of the attained network designs regarding the variations in the transportation costs.

We set the hub and gateway installation fixed costs (a^H and a^G), and the hub and gateway arc installation fixed costs (q^H and q^G) to 10,000.00, 100,000.00, 10.00, 1000.00, respectively, to assess how the optimal networks are affected by varying the scale economies or by varying unitary operational costs. In the first experiment, the domestic and international unitary operational costs (b^H and b^G) are both assumed as 1.00. Six different scenarios were constructed in which local (α^H) and global (α^G) scale economies were chosen within the set {0.2, 0.5, 0.8}, but having $\alpha^H \geq \alpha^G$ since it is expected that larger and

Table 13
Optimal network under different economies of scale values.

Instance	α^H	α^G	Time(s)	# Hubs	# Gateways	# Hub Arcs	# Gateway Arcs	PMDA	PMHA	PMGA
global-37	0.2	0.2	150	28	16	19	27	0.17	0.34	0.49
	0.5	0.2	171	28	19	10	35	0.28	0.20	0.52
	0.8	0.2	135	26	25	2	50	0.42	0.01	0.57
	0.5	0.5	163	28	20	16	39	0.28	0.22	0.50
	0.8	0.5	135	27	24	3	52	0.42	0.03	0.55
	0.8	0.8	147	27	23	15	46	0.43	0.06	0.51
global-43	0.2	0.2	694	32	17	25	27	0.19	0.37	0.44
	0.5	0.2	692	31	20	19	38	0.28	0.22	0.50
	0.8	0.2	676	28	25	4	51	0.44	0.02	0.54
	0.5	0.5	640	31	20	20	40	0.29	0.24	0.47
	0.8	0.5	646	28	24	6	54	0.45	0.03	0.52
	0.8	0.8	748	29	24	18	50	0.45	0.06	0.49
global-48	0.2	0.2	1447	35	17	34	28	0.22	0.41	0.37
	0.5	0.2	1137	34	22	23	44	0.32	0.25	0.43
	0.8	0.2	890	30	26	5	52	0.53	0.01	0.46
	0.5	0.5	1128	34	21	31	42	0.33	0.29	0.39
	0.8	0.5	873	31	24	11	54	0.53	0.04	0.43
	0.8	0.8	888	31	23	23	49	0.53	0.07	0.39

Table 14
Optimal network under different unitary operational costs.

Instance	b^H	b^G	Time(s)	# Hubs	# Gateways	# Hub Arcs	# Gateway Arcs	PMDA	PMHA	PMGA
global-37	0.5	0.5	325	28	16	19	27	0.05	0.47	0.49
	0.5	1.0	304	28	16	19	27	0.05	0.47	0.49
	0.5	1.5	243	28	16	19	27	0.05	0.47	0.49
	1.0	1.0	348	28	16	19	27	0.17	0.34	0.49
	1.0	1.5	242	28	16	19	27	0.17	0.34	0.49
	1.5	1.5	323	28	16	15	27	0.26	0.26	0.49
global-43	0.5	0.5	819	32	17	27	28	0.08	0.48	0.44
	0.5	1.0	691	32	17	27	28	0.08	0.48	0.44
	0.5	1.5	606	32	17	27	28	0.08	0.48	0.44
	1.0	1.0	661	32	17	25	27	0.19	0.37	0.44
	1.0	1.5	928	32	17	25	28	0.19	0.37	0.44
	1.5	1.5	838	32	17	21	28	0.27	0.29	0.44
global-48	0.5	0.5	1005	35	17	39	29	0.11	0.52	0.37
	0.5	1.0	744	35	17	38	28	0.11	0.52	0.37
	0.5	1.5	879	35	17	38	28	0.11	0.52	0.37
	1.0	1.0	1458	35	17	34	28	0.22	0.41	0.37
	1.0	1.5	925	35	17	34	28	0.22	0.41	0.37
	1.5	1.5	1162	35	17	30	28	0.30	0.33	0.37

more fuel efficient carriers are employed on gateway arcs. In the second experiment, to analyze the influence of unitary operational costs in the design of the network, six different scenarios were constructed in which local (b^H) and global (b^G) unitary operational costs were chosen within the set $\{0.5, 1.0, 1.5\}$, but having $b^H \le b^G$, since it is expected that international stopovers may involve performing customs, immigration and security checks. We adopted $\alpha^H = \alpha^G = 0.2$.

Tables 13 and 14 provide the results for variations in scale economies and unitary operational costs, respectively. The computational running times, the number of installed hubs and gateways, and hub and gateway arcs in the optimal solutions for the selected costs are reported. Tables 13 and 14 also indicate in percentage for each example three other performance measures, the Passenger Mile per Direct Arc (PMDA), Passenger Mile per Hub Arc (PMHA) and Passenger Mile per Gateway Arc (PMGA). Passenger Miles per arc type are calculated by summing the multiplication of the number of passengers that travelled through an arc by the arc length, as demonstrated in equations (48)–(50). By direct arc, we consider arc between two non-hub nodes or between a hub and a non-hub node.

$$PMDA = \sum_{(u,v)\in A^L} d_{uv} \sum_{(i,j)\in W} f_{uv}^{ij} \tag{48}$$

$$PMHA = \sum_{(u,v)\in A^H} d_{uv} \sum_{(i,j)\in W} f_{uv}^{ij} \tag{49}$$

$$PMGA = \sum_{(u,v)\in A^G} d_{uv} \sum_{(i,j)\in W} f_{uv}^{ij} \tag{50}$$

A detailed view of the results displayed in Table 13 shows a direct correlation between the optimal network and the relative difference in effectiveness for hubs and gateways. As the hub efficiency is lost, the investment priorities are re-routed through the international gateways. The same effect is observed on installation of hub and gateway arcs, when comparing to direct connections. In general, it is more attractive to select best global infrastructure and bring the international passengers close together using direct flights. Is important to recall that international demands can not be directly served, and therefore, as the gateways have their efficiency reduced, new attempts to improve the network cost structure are implemented, turning to hubs again as a cost-relieve device.

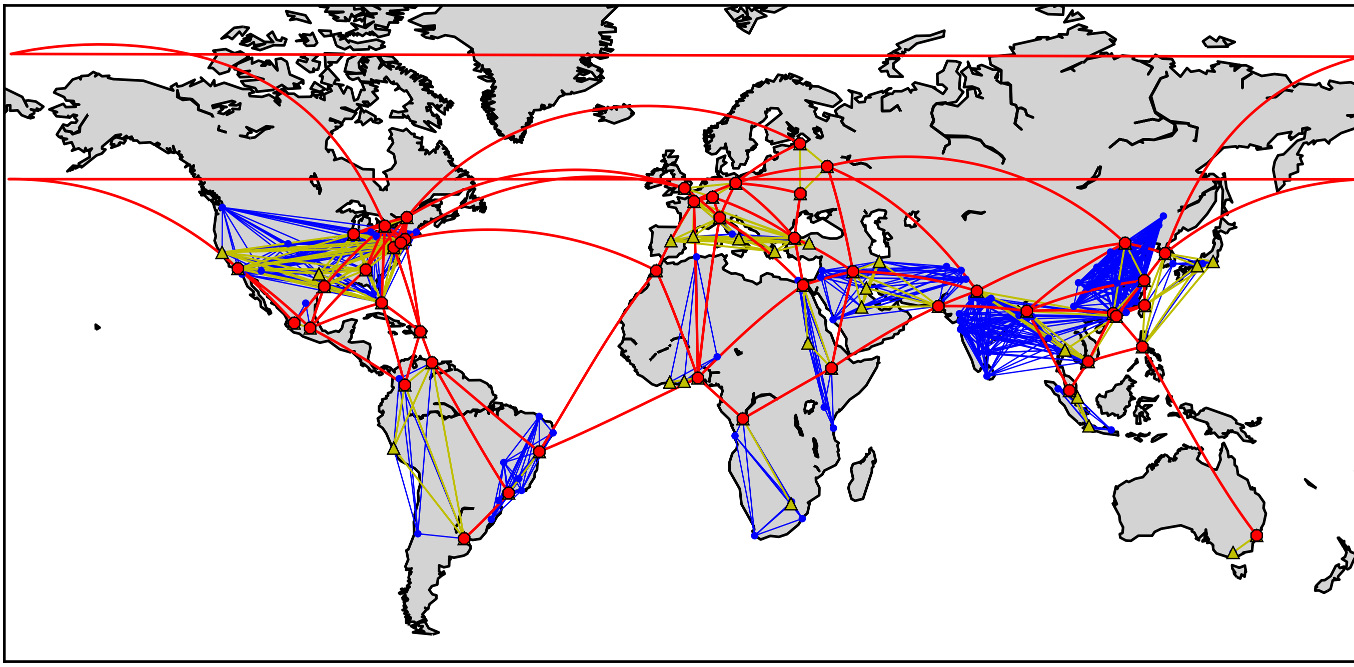


Fig. 2. Optimal air network for instance global-141 setting I.

Table 15
Total passenger traffic 2015.

Rank	Airport City/Country/Code	Solution
1	Atlanta GA, US (ATL)	Gateway
2	Beijing, CN (PEK)	Gateway
3	Dubai, AE (DXB)	Not in data
4	Chicago IL, US (ORD)	Gateway
5	Tokyo, JP (HND)	Hub
6	London, GB (LHR)	Gateway
7	Los Angeles CA, US (LAX)	Gateway
8	Hong Kong, HK (HKG)	Gateway
9	Paris, FR (CDG)	Gateway
10	Dallas/Fort Worth TX, US (DFW)	Hub
11	Istanbul, TR (IST)	Gateway
12	Frankfurt, DE (FRA)	Not in data
13	Shanghai, CN (PVG)	Gateway
14	Amsterdam, NL (AMS)	Not in data
15	New York NW, US (JFK)	Gateway
16	Singapore, SG (SIN)	Hub
17	Guangzhou, CN (CAN)	Gateway
18	Jakarta, ID (CGK)	Hub
19	Denver CO, US (DEN)	Not a candidate
20	Bangkok, TH (BKK)	Hub

Table 16
Total international passenger traffic 2015.

Rank	Airport City/Country/Code	Solution
1	Dubai, AE (DXB)	Not in data
2	London, GB (LHR)	Gateway
3	Hong Kong, HK (HKG)	Gateway
4	Paris, FR (CDG)	Gateway
5	Amsterdam, NL (AMS)	Not in data
6	Singapore, SG (SIN)	Hub
7	Frankfurt, DE (FRA)	Not in data
8	Incheon, KR (ICN)	Not in data
9	Bangkok, TH (BKK)	Hub
10	Istanbul, TR (IST)	Gateway
11	Taipei, TW (TPE)	Gateway
12	London, GB (LGW)	Gateway
13	Kuala Lumpur, MY (KUL)	Gateway
14	Madrid, ES (MAD)	Hub
15	Munich, DE (MUC)	Not in data
16	Doha, QA (DOH)	Not in data
17	Tokyo, JP (HND)	Hub
18	New York NW, US (JFK)	Gateway
19	Barcelona, ES (BCN)	Hub
20	Rome, IT (FCO)	Not a candidate

The important conclusion drawn here is that the intermediate network layer has its importance demonstrated specially for scenarios where a company is operating with old, not very effective international gateways like the ones available in on-development countries. Furthermore, ensuring the operational health and efficiency of both elements, hubs and gateways, is a task of capital importance to avoid unnecessary infrastructure mobilization. A careful planning of the installed capacity of these devices is advised, and an up-to-date tracking of good operational metrics to select the proper moment of capacity expansion as well.

As displayed in Table 14, it seems that the sensitivity to reasonable changes in unitary costs is not as high, regarding the structure of the optimal network. Only slight variations on the network topologies could be found, typically on the number of hub and gateway arcs. This is expected as the unitary costs were augmented using a flat profile to avoid introducing any bias towards any special location. As such, more subtle effects might be expected.

However, a detailed inspection of the metrics PMDA and PMHA show, as in the first experiment, that the local infrastructure or intermediate network layer has its importance reduced as the transportation costs in hubs and gateways get closer. Please, recall once more that no impact is expected in PMGA as there is no other way to cope with the global demand components except by going through the gateways. Once again, the direct flights become more and more attractive as the hubs display poor performance, in an effort to fast connect the international passengers. The hub layer infrastructure, to be important for relieving transportation costs, must be kept in good health, operating at good and safe overhead levels.

After the discussion of aforementioned results, the fundamental lesson is: a three layer network displaying specific and well defined roles for hubs and gateways may be of strategical value, provided that the transportation costs and economies of scale are kept under severe control. The differentiation between local and global infra-structure may be the key to relieve transportation costs when old, overloaded gateways are used, but to properly take advantage of the expected savings, it is required to keep the local hubs in good operational health. Otherwise, its is probably better dodge the costly stopovers at the hubs and feed the gateways using direct connections, returning to two layer network protocol.

5.2. An illustrative example

The problem proposed in this paper may act as a tool for providing insights into the current air network. In order to exemplify some practical benefits of the work developed here, the global-141 instance in setting I is chosen to be analyzed. Fig. 2 illustrates the network obtained for this instance. The top 20 busiest airports in the world considering total passenger traffic and total international passenger traffic in 2015 are demonstrated in Tables 15 and 16, respectively (ACI, 2016). The situation of the airports in the solution is highlighted in third column.

Comparing the real scenario with the network proposed, some questionings may arise. Although Atlanta airport was considered the busiest airport in 2015, should it really be an active gateway in the solution? This airport has a strategic location from a domestic point of view, being within a 2-h flight of 80% of the United States population (ACI, 2016). Nevertheless, ATL does not appear in international passenger traffic rank. In the same way, Beijing, Chicago, Los Angeles, Shanghai and Guangzhou airports appear as the top 20 busiest airports in Table 15 and also are active gateways in Fig. 2, but they are not present in Table 16. Should these airports be considered as gateways or as hubs? On the other hand, London, Hong Kong, Paris, Istanbul and New York airports are active gateways in our solution, being present in both ranks.

As the instances elaborated here contemplate the most populous cities, despite being well classified in both ranks, Dubai, Frankfurt and Amsterdam airports are not included in the instances. A similar situation is perceived with airports in Munich and Doha. They are in top 20 busiest international passenger traffic but they are not in the instance. In spite of being in data, Denver and Roma airports are not considered as gateway and hub candidates. This implies that instance creation is a point to be improved.

Dallas and Jakarta airports appear on the most-travelled airports in 2015 but they are not present in the rank when international passengers are considered. In the solution obtained for our model, they are active as hubs. On the contrary, although Tokyo, Singapore and Bangkok airports were considered as the most busiest airports of both total and only international passenger traffic, they are selected only as hubs in Fig. 2. Of course the already existing infrastructure of these airports must be considered, but considering those cities the best transshipment localities in Asia from an economic point of view is a matter of discussion. The optimal solution obtained for our model activated Ho Chi Minh City (Vietnam) airport as gateways in this region. Also in Asia, Incheon, Taipei and Kuala Lumpur airports are the world???s busiest in terms of international passengers and are selected as gateways in Fig. 2.

Despite being the fourteenth and the nineteenth world's busiest international passenger traficc airports, Madrid and Barcelona airports are proposed to be hubs in our solution. A similar question arises: Is Spain in a strategic place to concentrate international flows? The network drawn in Fig. 2 proposed Berlin and The Ruhr airports, in Germany, as gateways. Would Germany not be a more strategic country to concentrate international flights?

In our solution, Lagos, in Nigeria, and Casablanca, in Morocco, are proposed as gateways in Africa. Considering that Africa occupies a strategic position on the globe, these two countries would be strategic points for concentration of international flow?

In Brazil, Sao Paulo and Salvador airports are activated as gateways. We know that Rio de Janeiro airport concentrates much more flights than Salvador airport. From a geographic point of view, should be better to concentrate international flights in Salvador?

The costs considered in this study are not real, the demand utilized is based on the population of the city (we ignore the nearby cities) and the algorithms developed are only able to solve instances up to 141 nodes until optimality. Despite these limitations, the solution drawn here could provoke insights and questionings in the way of air companies prioritize airports.

6. Conclusion and future research

In this paper, we studied a gateway hub location problem, which is a three-level hub location problem, considering domestic and international flows for design a global air network. We have proposed exact methods to solve the problem, based on a generalized Benders decomposition. Two features that greatly speed up the method have been proposed: a separation of SECs when the BMP solution results in an infeasible network and a repair procedure which allows to generate Benders optimality cuts from unbounded dual subproblems. Computational experiments showed the effectiveness of our algorithms, which significantly improve the solution time of the general purpose solver when solving large instance sizes. As future research, the difference in performance between algorithms Alg-2-v2-I and Alg-3-v2 needs to be investigated further.

The formulation can tackle instances containing up to 141 airports, which is still far from real-life problem sizes. Moreover, the costs and the demand matrix considered here are not real data. Nevertheless, our approach offers important questions about the way current global air network is designed. Some efforts should be directed to improve the elaboration of the instances and also, other techniques can be studied to enable determining optimal solutions for larger instances.

References

Campbell, J.F., O'Kelly, M.E., 2012. Twenty-five years of hub location research. Transport. Sci. 46 (2), 153–169.
O'Kelly, M.E., 1987. A quadratic integer program for the location of interacting hub facilities. Eur. J. Oper. Res. 32 (3), 393–404.
ACI, 2016. Aci media Releases. Airports Council International Technical report.
Adler, N., 2005. Hub-spoke network choice under competition with an application to western europe. Transport. Sci. 39 (1), 58–72.
Adler, N., Smilowitz, K., 2007. Hub-and-spoke network alliances and mergers: price-location competition in the airline industry. Transp. Res. Part B Methodol. 41 (4), 394–409.
Ahuja, R., Magnanti, T., Orlin, J., 1993. Network Flows. Prentice-Hall, Upper Saddle River.
Alumur, S., Kara, B.Y., 2008. Network hub location problems: the state of the art. Eur. J. Oper. Res. 190 (1), 1–21.
Arslan, S., GreÅ, N., Ylmaz, H., 2011. A comparison of airline service expectations between passengers of domestic and international flights. Int. J. Soc. Sci. Humanity Stud. 3, 377–386.
Barnhart, C., Belobaba, P., Odoni, A.R., 2003. Applications of operations research in the air transport industry. Transport. Sci. 37 (4), 368–391.
Benders, J.F., 1962. Partitioning procedures for solving mixed-variables programming problems. Numer. Math. 4 (1), 238–252.
Birge, J.R., Louveaux, F.V., 1988. A multicut algorithm for two-stage stochastic linear programs. Eur. J. Oper. Res. 34 (3), 384–392.
Campbell, J.F., Ernst, A.T., Krishnamoorthy, M., 2002. Hub location problems. In: Drezner, Z., Hamacher, H. (Eds.), Facility Location: Application and Theory. Springer Berlin, pp. 373–407.
Catanzaro, D., Gourdin, E., Labbé, M., Özsoy, F.A., 2011. A branch-and-cut algorithm for the partitioning-hub location-routing problem. Comput. Oper. Res. 38 (2), 539–549.
Contreras, I., Cordeau, J.-F., Laporte, G., 2011. Benders decomposition for large-scale uncapacitated hub location. Oper. Res. 59 (6), 1477–1490.
Cordeau, J.-F., Soumis, F., Desrosiers, J., 2000. A benders decomposition approach for the locomotive and car assignment problem. Transport. Sci. 34 (2), 133–149.
Costa, A.M., 2005. A survey on benders decomposition applied to fixed-charge network design problems. Comput. Oper. Res. 32 (6), 1429–1450.
Farahani, R.Z., Hekmatfar, M., Arabani, A.B., Nikbakhsh, E., 2013. Hub location problems: a review of models, classification, solution techniques, and applications. Comput. Ind. Eng. 64 (4), 1096–1109.
Gelareh, S., Nickel, S., 2011. Hub location problems in transportation networks. Transport. Res. E Logist. Transport. Rev. 47 (6), 1092–1111.
Gelareh, S., Monemi, R.N., Nickel, S., 2015. Multi-period hub location problems in transportation. Transport. Res. E Logist. Transport. Rev. 75, 67–94.
Geoffrion, A.M., Graves, G.W., 1974. Multicommodity distribution system design by benders decomposition. Manag. Sci. 20 (5), 822–844.
Hansen, M., 1990. Airline competition in a hub-dominated environment: an application of

noncooperative game theory. Transp. Res. Part B Methodol. 24 (1), 27–43.
Hong, S., Harker, P.T., 1992. Air traffic network equilibrium: toward frequency, price and slot priority analysis. Transp. Res. Part B Methodol. 26 (4), 307–323.
Hsu, C.-I., Wen, Y.-H., 2003. Determining flight frequencies on an airline network with demandsupply interactions. Transport. Res. E Logist. Transport. Rev. 39 (6), 417–441.
IATA, 2015. 20 year Passenger Forecast. Technical Report. The International Air Transport Association.
Jr, J.T.B., 2012. A spatial analysis of fedex and ups: hubs, spokes, and network structure. J. Transport Geogr. 24, 419–431 Special Section on Theoretical Perspectives on Climate Change Mitigation in Transport.
Leung, J.M.Y., Magnanti, T.L., Singhal, V., 1990. Routing in point-to-point delivery systems: formulations and solution heuristics. Transport. Sci. 24 (4), 245–260.
Magnanti, T.L., Wong, R.T., 1981. Accelerating benders decomposition: algorithmic enhancement and model selection criteria. Oper. Res. 29 (3), 464–484.
Mao, L., Wu, X., Huang, Z., Tatem, A.J., 2015. Modeling monthly flows of global air travel passengers: an open-access data resource. J. Transport Geogr. 48, 52–60.
Mercier, A., Cordeau, J.-F., Soumis, F., 2005. A computational study of benders decomposition for the integrated aircraft routing and crew scheduling problem. Comput. Oper. Res. 32 (6), 1451–1476.
Papadakos, N., 2008. Practical enhancements to the magnantiwong method. Oper. Res. Lett. 36 (4), 444–449.
Preis, T., Huang, Z., Wu, X., Garcia, A.J., Fik, T.J., Tatem, A.J., 2013. An open-access modeled passenger flow matrix for the global air network in 2010. PLoS One 8 (5) e64317.
Sasaki, M., Furuta, T., Suzuki, A., 2009. Gateway location models. In: The Eighth International Symposium on Operations Research and its Applications (ISORA09). ORSC & APORC, pp. 356–363.
Schosser, M., Wittmer, A., 2015. Cost and revenue synergies in airline mergers: examining geographical differences. J. Air Transport. Manag. 47, 142–153.
Tarjan, R., 1972. Depth-first search and linear graph algorithms. SIAM J. Comput. 1 (2), 146–160.
Yaman, H., 2009. The hierarchical hub median problem with single assignment. Transp. Res. Part B Methodol. 43 (6), 643–658.