Title Page

Intermodal hub network design with generalized capacity constraints and non-synchronized train-truck operations

Mario José Basallo-Triana[a,*], Juan José Bravo-Bastidas[d], Ivan Contreras[b], Jean-François Cordeau[c], Carlos Julio Vidal-Holguín[d]

[a] *Universidad del Valle Sede Buga, School of Industrial Engineering, Carrera 13 No. 5-51, Buga, Colombia*
[b] *Concordia University and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, Quebec H3T 2A7, Canada*
[c] *HEC Montréal and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, Quebec H3T 2A7, Canada*
[d] *Universidad del Valle, School of Industrial Engineering, Calle 13 No. 100-00, Cali, Colombia*

Abstract

Current hub location models consider a single type of capacity constraint and assume that both import and export flows are processed at the same rate. This is not realistic in general, especially for intermodal hub networks. We posit that hub network design models should consider in more detail the characteristics of hub facilities and their main resources. To achieve a more realistic representation of hubs, we propose a hub location model that includes different types of hub resources and generalized capacity constraints. Generalized capacity constraints are necessary when the export and import processing rates are not equal or when there is a lack of synchronization between import and export vehicles, as is common in intermodal hub networks. We present and evaluate two mixed-integer linear programming formulations for the hub location problem with generalized capacity constraints. Formulations are evaluated under the well known AP data set and we also discuss the case study of a projected intermodal transportation system in Colombia. We conduct an analysis of the impact of train-truck synchronization on the rail-road hub network design. It was found that synchronization has an important impact on total costs affecting the hub network structure.

Keywords: Intermodal hub networks, hub location, generalized capacity constraints, synchronization, rail-road transshipment yard

Declarations of interest: none

*Corresponding author.
Email addresses: mario.basallo@correounivalle.edu.co (Mario José Basallo-Triana), juan.bravo@correounivalle.edu.co (Juan José Bravo-Bastidas), ivan.contreras@concordia.ca (Ivan Contreras), jean-francois.cordeau@hec.ca (Jean-François Cordeau), carlos.vidal@correounivalle.edu.co (Carlos Julio Vidal-Holguín)

Intermodal hub network design with generalized capacity constraints and non-synchronized train-truck operations

Mario José Basallo-Triana[a,*], Juan José Bravo-Bastidas[d], Ivan Contreras[b], Jean-François Cordeau[c], Carlos Julio Vidal-Holguín[d]

[a] *Universidad del Valle Sede Buga, School of Industrial Engineering, Carrera 13 No. 5-51, Buga, Colombia*
[b] *Concordia University and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, Quebec H3T 2A7, Canada*
[c] *HEC Montréal and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, Quebec H3T 2A7, Canada*
[d] *Universidad del Valle, School of Industrial Engineering, Calle 13 No. 100-00, Cali, Colombia*

Abstract

Current hub location models consider a single type of capacity constraint and assume that both import and export flows are processed at the same rate. This is not realistic in general, especially for intermodal hub networks. We posit that hub network design models should consider in more detail the characteristics of hub facilities and their main resources. To achieve a more realistic representation of hubs, we propose a hub location model that includes different types of hub resources and generalized capacity constraints. Generalized capacity constraints are necessary when the export and import processing rates are not equal or when there is a lack of synchronization between import and export vehicles, as is common in intermodal hub networks. We present and evaluate two mixed-integer linear programming formulations for the hub location problem with generalized capacity constraints. Formulations are evaluated under the well known AP data set and we also discuss the case study of a projected intermodal transportation system in Colombia. We conduct an analysis of the impact of train-truck synchronization on the rail-road hub network design. It was found that synchronization has an important impact on total costs affecting the hub network structure.

Keywords: Intermodal hub networks, hub location, generalized capacity constraints, synchronization, rail-road transshipment yard

*Corresponding author.
Email addresses: mario.basallo@correounivalle.edu.co (Mario José Basallo-Triana), juan.bravo@correounivalle.edu.co (Juan José Bravo-Bastidas), ivan.contreras@concordia.ca (Ivan Contreras), jean-francois.cordeau@hec.ca (Jean-François Cordeau), carlos.vidal@correounivalle.edu.co (Carlos Julio Vidal-Holguín)

1. Introduction

Hub networks are used for the routing of flows between different origin and destination nodes. In this setting, a set of hub facilities allows sorting and consolidation of flows coming from low-capacity vehicles. Then, high-capacity vehicles perform consolidated shipments between hubs. Finally, the traffic is deconsolidated and distributed to the destination nodes. The practical attractiveness of hub networks comes from the reduction in transportation costs due to economies of scale in inter-hub transportation along with traffic reduction, which is a consequence of consolidation at hubs. Hub location models aim to determine the location of the hubs and the routing of flows from origin nodes to destination nodes passing through a set of hubs. Reviews on hub location modeling include Alumur & Kara (2008), Campbell & O'Kelly (2012), Contreras & O'Kelly (2019) and Alumur et al. (2021). A review of hub location modeling in the context of intermodal transportation is presented by Basallo-Triana et al. (2021).

In this paper we present a new hub location problem motivated by the modeling of intermodal rail-road transportation networks. Rail-road networks are designed to satisfy the transport demand of many-to-many distribution systems. Combined rail-road transportation has attracted a lot of attention because of its ability to attain lower transport costs due to the consolidation of large freight volumes at transshipment facilities and their transportation by train over long distances (Janic, 2007; Ishfaq & Sox, 2010). Rail-road networks are also attractive because of the mitigation of the environmental impact that freight transportation poses to our planet (Steadieseifi et al., 2014). The design and operation of rail-road networks can be done from a hub location perspective (Campbell & O'Kelly, 2012; Crainic & Kim, 2007).

A hub or transshipment facility is a key element that determines the competitiveness of the rail-road hub network (Kreutzberger & Konings, 2016; Boysen et al., 2013). This is because transshipment operations at hubs may generate inefficiencies affecting the performance of the entire transportation system (Saeedi et al., 2019; Boysen et al., 2013). It is noted, however, that rail-road transshipment facilities have characteristics which are different from other types of hub facilities. On the one hand, train and truck arrivals are not necessarily synchronized, implying that a portion of the freight moved through each hub must be temporarily stored. This makes it necessary to consider the storage space availability as a limiting resource which also affects the performance of the transshipment equipment. On the other hand, the processing rate of the flow that comes to a hub from non-hub nodes, which we refer to as export flow, is not necessarily the same as the processing rate of the flow that comes to a hub from other hub nodes, which we refer to as import

flow. In intermodal transportation, import operations tend to be more time consuming than export operations (Crainic & Kim, 2007).

The above considerations suggest that hubs should be modelled as more complex facilities than what is usually done in existing hub location models. In this sense, the hub is a facility operating with multiple resources having limited capacity. Also, the processing rate of import and export flows is not necessarily equal. This leads us to introduce *generalized capacity constraints*, which are defined for different hub resources and consider the imbalance in the processing rates of import and export flows. This concept is a generalization of traditional capacity constraints in hub location and allows greater flexibility in hub design. For example, the hub network design model may consider multiple hub configurations or capacity levels with different transshipment equipment and storage space characteristics.

Because rail-road transshipment yards are expensive facilities performing time-intensive and resource-consuming operations (Wiegmans & Behdani, 2018), the design of intermodal hub networks should consider, with a greater level of detail, the type of operations and resources of such yards, as well as synchronization aspects. In practice, inappropriate terminal planning and network design have led to the collapse of entire intermodal hub networks, as was the case of the innovative transshipment yard Mainhub Antwerp within the NARCON network in Europe (Kreutzberger & Konings, 2016).

In this paper, we present a path-based formulation of a multiple allocation capacitated hub location model with multiple capacity levels or hub configurations along with multiple hub resources and with direct transportation decisions. Economies of scale are modeled by using a constant discount factor in the inter-hub transportation costs. We propose two mixed-integer linear programming reformulations of generalized capacity constraints.

The remainder of this paper is structured as follows. Section 2 presents a literature review. Section 3 describes the hub location model with generalized capacity constraints and two mixed-integer linear programming formulations. Section 4 shows the application of the hub location model with generalized capacity constraints to the design of rail-road intermodal hub networks with non-synchronized train-truck operations. Section 5 evaluates the performance of the formulations, analyzes the impact of synchronization on the performance of the network, and discusses a case study. Finally, Section 6 concludes the paper.

2. Literature review

Capacitated hub location problems were first formulated for a single capacity level (Campbell, 1994). Most of the early works defined capacity constraints only for the flow that comes from non-hub nodes (Ebery et al., 2000; Boland et al., 2004; Marín, 2005). The reason for this comes from the postal delivery applications, where the sorting operations are required only for the flow that comes from non-hub nodes (Ebery et al., 2000). There are also computational advantages in defining capacity constraints in this way. For example, this allowed Contreras et al. (2012) to solve primal sub-problems in a Benders decomposition approach as transportation problems, which makes it possible to solve instances with up to 300 nodes. However, defining capacity constraints by considering only the flow that comes from non-hub nodes is not realistic in general. Note that the absence of sorting operations for the flow that comes from hub nodes does not imply that such flow is processed at an infinity processing rate. Taherkhani et al. (2020) propose a Benders decomposition algorithm for a path-based hub location model with multiple capacity levels including capacity constraints for the flow that comes from hub and non-hub nodes. The proposed algorithm can solve instances with up to 500 nodes. Their formulation assumes the same processing rate for import and export flows.

Capacitated hub location models focus on the definition of multiple capacity levels. The common approach is to use binary variables for selecting the hub capacity from a set of capacity levels (Alumur et al., 2016; Azizi et al., 2018; Marufuzzaman et al., 2014; Zhalechian et al., 2017a,b). Some authors define capacity constraints for different modes of transport (Serper & Alumur, 2016; Mohammadi et al., 2019). The use of continuous variables for defining hub capacities is less common in practice (Ghaffari-Nasab et al., 2015; Fotuhi & Huynh, 2015). Besides, Kahag et al. (2019) use integer variables for selecting the number of servers installed in a hub considering an M/M/c queuing model.

The literature on rail-road hub network design includes Arnold et al. (2004), who propose a hub location model with a formulation that is closely linked to a fixed charge multi-commodity network design model. Limbourg & Jourquin (2009) propose a single allocation p-hub location model with a flow-based formulation for locating rail-road transshipment yards. Meanwhile, Racunica & Wynter (2005) and Ishfaq & Sox (2010) introduce concave transportation cost functions to account for the economies of scale in inter-hub transportation. To handle the non-linearity of the cost function, they propose a piece-wise linear approximation (see also Wang & Meng, 2017). Both groups of authors use a path-based formulation. Meng & Wang (2011) use a translog cost function in a

scenario with multiple types of containers to reflect economies of scale along with transitions to diseconomies of scale at hubs.

Other works consider single allocation p-hub location models with service time constraints (Ishfaq & Sox, 2010, 2011, 2012; Yang et al., 2016; Wang et al., 2018). All of them formulate the problem with a path-based formulation. The work by Teye et al. (2018) (see also Teye et al., 2017) propose the entropy maximizing approach to increase the terminal usage in rail-road hub networks.

Some researchers have introduced rail-road hub network design models considering uncertainty in data. Uncertainty in transport cost and transport time is considered by Yang et al. (2016) and Wang et al. (2018). For their part, Li et al. (2017) and Merakli & Yaman (2016) assume uncertainty in transport demand. In their work, Ishfaq & Sox (2011) consider that hub operations comprise three stages: the unloading operation, the freight consolidation into batches or freight deconsolidation, and, finally, the loading operation. Each hub facility is modeled using a GI/G/1 queue system. Then, service time constraints include the transport and waiting time at hubs due to congestion.

Capacitated hub location models are uncommon in the intermodal rail-road hub network literature. In this case, Sörensen & Vanovermeire (2013) consider a capacitated multiple allocation p-hub location model with a single capacity level and direct transportation decisions. More efficient reformulations of the capacitated model than the one by Sörensen & Vanovermeire (2013) are provided by Lin et al. (2014) and Lin & Lin (2016).

The contributions of the present paper are the development of a modeling approach for hub location with more realistic capacity constraints by including multiple hub resources along with unbalanced processing rates for import and export operations. Two formulations are provided for generalized capacity constraints. We also provide a modeling framework for considering non-synchronized train-truck operations in transshipment yards from a stochastic point of view. The consideration of non-synchronized operations shows the necessity of using generalized capacity constraints in rail-road transportation networks. To the best of our knowledge, these features have not been previously studied in the literature.

3. Intermodal hub location with generalized capacity constraints

Let $G = (N, A)$ be a complete digraph, where N is the set of nodes and A is the set of arcs. A hub arc a is defined as the ordered set (a_1, a_2), where $a_1 \in N$ and $a_2 \in N$ are hub nodes. Let K be the set of commodities or origin-destination (OD) pairs whose origin and destination nodes belong

to N, which is defined for strictly different OD nodes as $K = \{(i,j) \in N \times N \mid w_{(i,j)} > 0, i \neq j\}$, where w_k is the annual transport demand for commodity $k \in K$ that include nodes i and j. Let R be the set of hub resources, including possibly the transshipment equipment, work force, storage space availability, and others. We define the set Q as the set of different hub configurations. A hub configuration corresponds to a specific hub design with a given installed capacity for each resource $r \in R$.

Let $C_{ka} = C_{k,(a_1,a_2)}$ be the transportation cost for commodity $k \in K$ through hub arc $a = (a_1, a_2) \in A$. This cost parameter is defined as $C_{ka} = C_{(i,j),(a_1,a_2)} = w_k(\chi c_{ia_1} + \tau c_{a_1a_2} + \delta c_{a_2j})$, where c_{ij} is the travel distance between nodes $i \in N$ and $j \in N$. The quantity C_{ka} is also referred to as the intermodal transport cost. Here, χ is the unit collection cost, τ is the unit transfer cost, and δ is the unit distribution cost. To reflect economies of scale we assume that $\tau < \chi$ and $\tau < \delta$.

Let C_k be the direct transportation cost for commodity k, which is defined as $C_k = w_k \rho c_k$. We denote by c_k the travel distance for satisfying the demand of commodity k, i.e., the distance between the origin node i and destination node j, where $k = (i,j)$. For its part, ρ is the unit direct transportation costs. It is reasonable to assume that $\rho \geq \chi$ and $\rho \geq \delta$. Let f_{hq} be the annualized fixed installation cost of hub configuration $q \in Q$ at node $h \in N$.

The rail-road intermodal hub location problem has characteristics that differ from other hub location models. Given that the purpose of the hub equipment is to perform a transshipment from one mode of transport to another and that yard equipment is an expensive resource, only two-hub paths are allowed. Single-hub paths are disregarded due to unnecessarily increasing the use of expensive yard resources. We further define the set A_k as the set of hub arcs that can be used to satisfy the demand of commodity $k \in K$, which have the least cost path between two distinct hub nodes (see Taherkhani et al., 2020) and for which the intermodal transport option is cheaper than the direct transport option:

$$A_k = \{a \in A \mid a_1 \neq a_2, C_{k(a_1,a_2)} < C_{k(a_2,a_1)}, C_{k(a_1,a_2)} < C_k\}. \tag{1}$$

In some situations it might be necessary to satisfy service time requirements. This could be done during a preprocessing step. Let t_{ij} be the transport time between nodes i and j using direct transportation and let τ_{ij} be the maximum time for satisfying the demand between the pair of nodes i and j. The time for the intermodal transport option is defined as $T_{ka} = t_{ia_1} + \gamma t_{a_1a_2} + t_{a_2j}$, where γ is a factor used for computing the time for the transportation between hubs. In rail-road transportation it is expected that $\gamma > 1$. Therefore, the set A_k can by redefined so that service

time requirements are satisfied:

$$A_k = \left\{ a \in A \mid a_1 \neq a_2, C_{k(a_1,a_2)} < C_{k(a_2,a_1)}, C_{k(a_1,a_2)} < C_k, T_{k(a_1,a_2)} < \tau_k \right\}.$$

Let $x_{ka} \geq 0$ be a continuous decision variable used to represent the fraction of the flow of commodity $k \in K$ that is routed using the intermodal transportation option using hub arc $a \in A_k$.

In intermodal transportation, we distinguish two main types of flows entering a hub, therefore two main types of transshipment operations. The *export flow* is the flow that comes to a hub from non-hub nodes by means of low-capacity vehicles e.g., trucks. In contrast, the *import flow* is the flow that comes to a hub from other hubs by means of high-capacity vehicles e.g., trains or ships. For hub $h \in N$, the export flow is denoted as λ_h^{e} and the import flow is denoted as λ_h^{i}. These flows can be defined in terms of the decision variables x_{ka} as follows:

$$\lambda_h^{\mathrm{e}} = \sum_{k \in K} \sum_{\substack{a \in A_k, \\ a_1 = h}} w_k x_{ka}, \qquad \forall h \in N, \tag{2}$$

$$\lambda_h^{\mathrm{i}} = \sum_{k \in K} \sum_{\substack{a \in A_k, \\ a_2 = h}} w_k x_{ka}, \qquad \forall h \in N. \tag{3}$$

Let $L_{hq}^r > 0$ be the installed capacity at hub $h \in N$ for resource $r \in R$ and hub configuration $q \in Q$. Let also $a_{hq}^r > 0$ be the unit capacity consumption for export operations, and $b_{hq}^r > 0$ be the unit capacity consumption for import operations related to resource $r \in R$ and hub configuration $q \in Q$, respectively. Consider for example that $a_{hq}^r = 0.1$ hours per unit, then, each export operation consumes 0.1 hours of the installed capacity. If the installed capacity is $L_{hq}^r = 3{,}000$ hours per year then at most 30,000 export operations can be carried out before exhausting the installed capacity. In general, the total capacity consumption for resource $r \in R$ at hub $h \in N$ is $a_{hq}^r \lambda_h^{\mathrm{e}} + b_{hq}^r \lambda_h^{\mathrm{i}}$, and the corresponding generalized capacity constraint is $a_{hq}^r \lambda_h^{\mathrm{e}} + b_{hq}^r \lambda_h^{\mathrm{i}} \leq L_{hq}^r$. Note that the unit of measurement can be anything and does not have to be hours.

There exists a more involved interpretation of capacity related parameters that considers the inherent uncertainty in data and in which the hub is viewed as a queuing system serving two different kinds of units, i.e., import and export units. Assume that the number of servers in hub $h \in N$ for resource $r \in R$ and hub configuration $q \in Q$ is L_{hq}^r. The average processing time of an export unit is a_{hq}^r and the average processing time for an import unit is b_{hq}^r. Note that the probability distribution of these two parameters does not matter as long as the variance of such distribution is finite. If a hub is established at node $h \in N$ with configuration $q \in Q$, a condition

that guarantees that the number of units in the system does not grow infinitely is (Shortle et al., 2018, p. 174):

$$a^r_{hq}\lambda^{\mathrm{e}}_h + b_{hq}\lambda^{\mathrm{i}}_h < L^r_{hq}. \tag{4}$$

This is called the stability condition of the system, which can be interpreted as a sort of capacity constraint. The left hand side of (4) is referred to as the total traffic intensity and the quantity $(a^r_{hq}\lambda^{\mathrm{e}}_h + b^r_{hq}\lambda^{\mathrm{i}}_h)/L^r_{hq}$ is the expected server (hub) utilization.

In general, the unit capacity consumption is different for export and import flows. For the intermodal transportation case, import operations tend to have a higher capacity consumption when compared to export operations. This is due to the nature of the storage systems in intermodal facilities. Most seaports and rail-road terminals operate with grounded storage and stacking operations. In those systems, once the high-capacity vehicle is in the transshipment area, export operations are carried out efficiently following a load plan that avoids rehandlings in the storage yard. However, given that the arrivals of low-capacity vehicles (trucks) occur randomly, it is more difficult to prevent rehandlings for import operations. Thus, most import operations require higher handling efforts compared to export operations, see Crainic & Kim (2007) and Basallo-Triana et al. (2022).

Dominant resources

In practice, there might be resources that, for a given hub configuration $q \in Q$, produce redundant capacity constraints. We refer to such resources as non-dominant and they should not be considered in the implementation of the model. Figure 1 shows a representation of generalized capacity constraints, which is helpful for understanding the logic behind the identification of dominant resources. In the figure, dominant resources include [1], [2] and [4]. These resources determine the flow patterns by defining a convex feasible region for the export and import flow at each facility. Let $R_q \subseteq R$ be the set of dominant resources for hub configuration $q \in Q$.

3.1. Big-M formulation

There are several ways to model generalized capacity constraints. Some formulations preserve the size of the problem in terms of the number of continuous variables. Such formulations contain big-M parameters and are also referred as *big-M formulations*. One disadvantage of big-M formulations is the loss of formulation strength (Vielma, 2015).

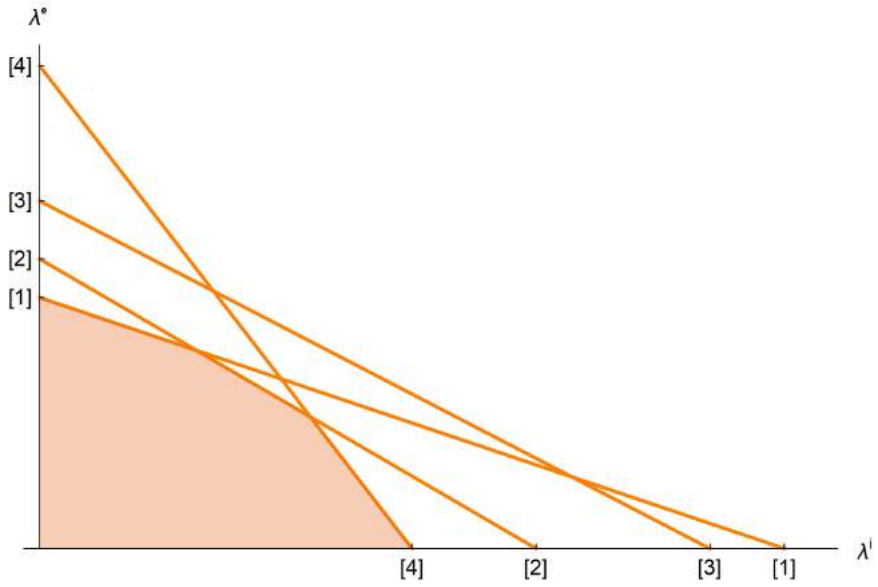

Figure 1: Nature of generalized capacity constraints. Dominant resources: [1], [2], [4]. Non-dominant resource: [3].

The big-M formulation for the hub location problem with generalized capacity constraints considered in this paper is referred to as `BM`. Let z_{hq} be a binary variable with a value of 1 if a hub is established at node $h \in N$ with hub configuration $q \in Q$; otherwise 0. The `BM` formulation is

$$\texttt{BM}: \quad \text{minimize} \quad \sum_{h\in N}\sum_{q\in Q} f_{hq} z_{hq} + \sum_{k\in K}\sum_{a\in A_k} C_{ka} x_{ka} + \sum_{k\in K} C_k \left(1 - \sum_{a\in A_k} x_{ka}\right) \tag{5}$$

$$\text{subject to} \quad \sum_{a\in A_k} x_{ka} \leq 1, \qquad \forall k \in K, \tag{6}$$

$$\sum_{\substack{a\in A_k,\\ h\in a}} x_{ka} \leq \sum_{q\in Q} z_{hq}, \qquad \forall h \in N, k \in K, \tag{7}$$

$$a^r_{hq} \sum_{k\in K}\sum_{\substack{a\in A_k,\\ h=a_1}} w_k x_{ka} + b^r_{hq} \sum_{k\in K}\sum_{\substack{a\in A_k,\\ h=a_2}} w_k x_{ka} \leq \sum_{l\in Q} M^r_{hlq} z_{hl},$$

$$\forall h \in N, r \in R, q \in Q, \tag{8}$$

$$\sum_{q\in Q} z_{hq} \leq 1, \qquad \forall h \in N, \tag{9}$$

$$z_{hq} \in \{0,1\}, \qquad \forall h \in N, q \in Q, \tag{10}$$

$$0 \leq x_{ka} \leq 1, \qquad \forall k \in K, a \in A_k. \tag{11}$$

The objective function (5) minimizes the total annual cost. This includes the annualized investment in hub facilities given in the first term, the annual intermodal transportation cost given in the second term, and the annual direct or truck-only transportation cost given in the third term.

Constraints (6) guarantee that the demand of each commodity is satisfied using the intermodal transport option, the direct or truck-only transport option or any combination of the previous options. Constraints (7) prohibit commodities from being routed through closed hubs. Constraints (8) are generalized capacity constraints, where M^r_{hlq} is a big-M parameter. These constraints have the form $a^r_{hq}\lambda^{\mathrm{e}}_h + b^r_{hq}\lambda^{\mathrm{i}}_h \leq \sum_{l\in Q} M^r_{hlq} z_{hl}$, where parameter M^r_{hlq} is carefully chosen so that these inequalities are valid. Constraints (9) guarantee that at most one hub configuration can be established at a given node. Constraints (10) and (11) establish the domain of the decision variables.

Note that objective function is simplified as

$$\sum_{h\in N}\sum_{q\in Q} f_{hq} z_{hq} - \sum_{k\in K}\sum_{a\in A_k} (C_k - C_{ka})\, x_{ka} + \sum_{k\in K} C_k. \tag{12}$$

The term $C_k - C_{ka}$ is the saving of using intermodal transportation on path a for commodity k instead the direct transportation, which, by definition of A_k (see equation (1)), must be a positive quantity. This means that intermodal transportation is an attractive option if there exists an inter-hub network, with at least two hubs, in which the annual investment cost on hub facilities is less than the total intermodal transport savings. The following result provides the tightest possible values for the M^r_{hlq} parameters that lead to valid generalized capacity constraints.

Proposition 1. *Valid values for the big-M parameters are*

$$M^r_{hlq} = \begin{cases} L^r_{hq}, & \text{if } l = q, \\ \max_{\lambda^e,\lambda^i} \left\{ a^r_{hq}\lambda^e + b^r_{hq}\lambda^i \mid a^j_{hl}\lambda^e + b^j_{hl}\lambda^i \leq L^r_{hl}, \lambda^e \geq 0, \lambda^i \geq 0, \forall j \in R_l \right\}, & \text{if } l \neq q. \end{cases} \tag{13}$$

Proof. Consider that $z_{hl} = 1$ and $l = q$, the feasible set given by generalized capacity constraints for hub configuration q is $\mathcal{R}_l = \{a^r_{hl}\lambda^{\mathrm{e}}_h + b^r_{hl}\lambda^{\mathrm{i}}_h \leq L^r_{hl}, \lambda^{\mathrm{e}}_h \geq 0, \lambda^{\mathrm{i}}_h \geq 0, \forall r \in R_l\}$. Such constraints determine the flow pattern through hub $h \in N$. Given that $z_{hl} = 1$, any other capacity constraint different from the set $\mathcal{R}_l$ cannot alter the flow pattern and must be redundant.

Now consider the case $z_{hl} = 1$ and $l \neq q$, generalized capacity constraints for hub configuration q are $a^r_{hq}\lambda^{\mathrm{e}}_h + b^r_{hq}\lambda^{\mathrm{i}}_h \leq M^r_{hlq} = \max_{\lambda^{\mathrm{e}},\lambda^{\mathrm{i}}} \left\{ a^r_{hq}\lambda^{\mathrm{e}} + b^r_{hq}\lambda^{\mathrm{i}} \mid a^j_{hl}\lambda^{\mathrm{e}} + b^j_{hl}\lambda^{\mathrm{i}} \leq L^r_{hl}, \lambda^{\mathrm{e}} \geq 0, \lambda^{\mathrm{i}} \geq 0, \forall j \in R_l \right\}$. Since M^r_{hlq} is an extremum value for region $\mathcal{R}_l$, the previous constraints do not affect the flow patterns through hub h and hence are redundant. □

3.2. Multiple choice formulation (MC)

The multiple choice (MC) formulation is an extended formulation because it introduces auxiliary continuous variables, which helps to improve the characterization of the feasible region in a higher dimensional space. The main property of the MC formulation is that the linear programming relaxation related to generalized capacity constraints is tight (Vielma, 2015).

We refer to this formulation as MC (see Balas, 1998; Jeroslow & Lowe, 1983). Let $\lambda^{\mathrm{e}}_{hq} \geq 0$ and $\lambda^{\mathrm{i}}_{hq} \geq 0$, for $h \in N$ and $q \in Q$, be new non-negative continuous variables. The MC formulation is

$$\begin{aligned}
\texttt{MC}: \quad \text{minimize} \quad & \sum_{h\in N}\sum_{q\in Q} f_{hq} z_{hq} - \sum_{k\in K}\sum_{a\in A_k} (C_k - C_{ka})\, x_{ka} + \sum_{k\in K} C_k. \\
\text{subject to} \quad & (6), (7), (9) - (11), \\
& \sum_{k\in K}\sum_{\substack{a\in A_k,\\ h=a_1}} w_k x_{ka} = \sum_{q\in Q} \lambda^{\mathrm{e}}_{hq}, \qquad \forall h \in N, && (14)\\
& \sum_{k\in K}\sum_{\substack{a\in A_k,\\ h=a_2}} w_k x_{ka} = \sum_{q\in Q} \lambda^{\mathrm{i}}_{hq}, \qquad \forall h \in N, && (15)\\
& a^r_{hq}\lambda^{\mathrm{e}}_{hq} + b^r_{hq}\lambda^{\mathrm{i}}_{hq} \leq L^r_{hq} z_{hq}, \qquad \forall r \in R_q, h \in N, q \in Q, && (16)\\
& \lambda^{\mathrm{e}}_{hq}, \lambda^{\mathrm{i}}_{hq} \geq 0, \qquad \forall h \in N, q \in Q. && (17)
\end{aligned}$$

This basic idea behind the MC formulation is to create $|Q|$ copies of each variable λ^{e}_h and λ^{i}_h in order to model separately the associated generalized capacity constraints. Alternatively, it is possible to create $|Q|$ copies of flow variables by defining them as x^q_{ak}, for $q \in Q$. It should be noted that this leads to a considerable increase in the size of the problem. This strategy was used by Najy & Diabat (2020) for the modeling of a different class of disjunctive constraints involving piece-wise linear approximations of congestion and transportation cost functions. The strategy may produce advantages when using specialized solution algorithms such as Benders decomposition.

4. Application to rail-road hub network design with non-synchronized train-truck operations

We next describe the application of generalized capacity constraints to a rail-road hub network design problem, which is a result of the consideration of synchronization aspects. Intermodal rail-road transportation is a key part of inland transportation with economic, environmental and low congestion advantages, being of high interest in practice. We first describe the characteristics of rail-road transshipment yards. We then develop generalized capacity constraints considering the

transshipment equipment and the storage space availability as main resources. To this end, we present a thorough analysis of train-truck synchronization.

4.1. Description of a rail-road transshipment yard

A rail-road transshipment yard is equipped with rail-mounted gantry cranes for performing transshipment operations between trains and trucks, with a typical structure shown in Figure 2. The yard comprises a set of transshipment tracks for unloading (loading) units from (to) trains. Parking and driving lanes facilitate truck exchange operations. There is also a storage area for stacking intermodal transport units (ITUs). Typically, the rail tracks, truck lanes, and storage area are located between the crane legs.

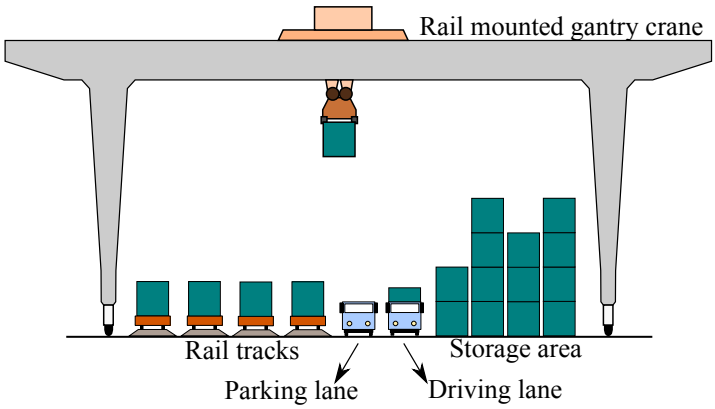

Figure 2: Illustration of a rail-road transshipment yard. Source: Basallo-Triana et al. (2022).

We distinguish two kinds of transshipment operations. A *direct transshipment* occurs when the ITU is transshipped directly from the train to the truck or vice versa, without intermediate storage. In contrast, an *indirect transshipment* occurs when a direct transshipment cannot be performed, and the ITU needs to be located temporarily in the storage area. Indirect operations are time-consuming tasks requiring multiple handling efforts by cranes. The main cause of indirect operations is the lack of synchronization between trains and trucks.

There are two types of flows entering a transshipment yard. The *export flow* refers to ITUs that arrive by truck from some origin node and leave the yard by train. The *import flow* refers to ITUs that arrive by train from other transshipment yards and leave the yard by truck. We note that the amounts of direct and indirect transshipments and export and import operations have different impacts on yard performance, as we will show later.

Gantry cranes perform the following series of operations:

- Direct transshipment (dt): including wagon-to-truck and truck-to-wagon operations,
- Wagon-to-storage operations (ws),
- Storage-to-wagon operations (sw),
- Truck-to-storage operations (ts),
- Storage-to-truck operations (st), including possible rehandling operations.

Rehandling operations occur when an ITU needs to be extracted from a stack, and there are other ITUs above it, so it is necessary to relocate such units. Rehandlings are expected only in storage-to-truck operations because of the random nature of truck arrivals. These operations are not necessary for storage-to-wagon operations due to a planned train loading sequence (Crainic & Kim, 2007). Basallo-Triana et al. (2022) derived analytical expressions for computing the expected processing times of the different types of operations described above. We use these expressions in this work to account for the capacity of the handling equipment and the use of storage space.

4.2. Rail-road hub network design assuming complete train-truck synchronization

The rail-road hub network design problem consists of locating a set of rail-road transshipment yards or hubs, determining the transportation pattern of ITUs for each OD node, and for selecting the yard configuration or design of each facility. When train and truck arrivals are synchronized, ITUs are directly transshipped from one mode of transport to the other without requiring intermediate storage. The expected transshipment time from a train to a truck is the same as the expected transshipment time from a truck to a train. This leads to the simpler *hub location problem with traditional capacity constraints* (`HLPTCC`) (see Taherkhani et al., 2020):

$$
\begin{aligned}
\texttt{HLPTCC}: \quad & \text{minimize} && (5) \\
& \text{subject to} && (6), (7), (9) - (11), \\
& && \sum_{k \in K} \sum_{\substack{a \in A_k, \\ h=a}} w_k x_{ka} \leq \sum_{q \in Q} \frac{L^1_{hq}}{T^{\mathrm{dt}}_q} z_{hq}, \qquad \forall h \in N, \qquad (18)
\end{aligned}
$$

where T^{dt}_q is the average cycle time for a direct transshipment operation in yard configuration $q \in Q$, then $1/T^{\mathrm{dt}}_q$ is the corresponding average processing rate (see Basallo-Triana et al., 2022). It is important to note that w_k and T^{dt}_q must be given in the same time units. L^1_{hq} is the number of rail-mounted gantry cranes for yard configuration q. Constraints (18) are referred to as traditional

capacity constraints limiting the total (export and import) transport flow through the yard (see Taherkhani et al., 2020). Traditional capacity constraints assume that the processing or consumption rates are the same for both export and import operations. We will show that this is not correct in general.

4.3. An approximation to the rail-road hub network design considering non-synchronized train-truck operations

The modeling of rail-road hub networks with non-synchronization is by far more challenging than the synchronized case. In non-synchronized operations, yard performance depends on the interactions with the storage area, which is a limited resource. These interactions reduce the capacity of the handling equipment compared to the synchronized case. Due to the presence of rehandlings in indirect import operations, export operations tend to be more efficient or consume less resources. In this sense, there are two types of resources of interest: the rail-mounted gantry cranes and the storage space availability.

We present a simple treatment of non synchronization aimed to maintain the tractability of the hub location model and preserve its linearity without introducing additional decision variables. Our formulation takes a stochastic point of view of synchronization in which the level of synchronization is given by the distributional characteristics of the ITU dwell times and truck arrival rates, with respect to the train service time. This differs from other approaches that take a deterministic point of view of synchronization. Such approaches may consider the arrival and departure time of vehicles as decision variables and the level of synchronization is given by the degree of overlap in the arrival and departure date intervals for two distinct vehicles. See (Alumur, 2009, p. 24–44) for a related discussion on synchronization.

To develop a quantitative description of the terminal in this context, we consider the following assumptions:

1. The expected portion of direct transshipments for export and import ITUs are known in advance.
2. Trains are served in train bundles or batches with high utilization of rail tracks. This implies that several trains are served simultaneously. This form of operation increases terminal efficiency since crane utilization might be better balanced in batch operations (Boysen et al., 2013; Rotter, 2004).

3. The truck arrival pattern follows a Poisson process. This assumption is commonly adopted in planning models for rail-road terminals (García & García, 2012; Rizzoli et al., 2002; Corry & Kozan, 2006).
4. ITU dwell times are exponentially distributed. In this work, the ITU dwell time must be understood as the time an ITU spends in the terminal. This includes the time the container spends in a rail wagon and, if required, the time for temporary storage.

Let r_h^{e} and r_h^{i} be the fractions of direct transshipments for export and import ITUs, respectively, in node $h \in N$, which are assumed to be known in advance. Let λ_h^{e} and λ_h^{i} be the corresponding average ITU arrival rates to hub h. The main ITU flows through the different areas in the yard are depicted in Figure 3.

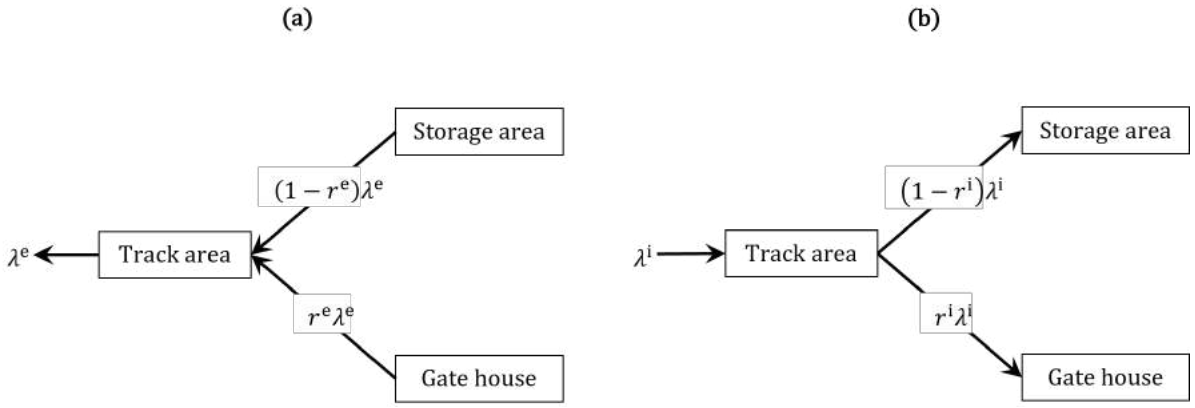

Figure 3: **(a)** Export flows. **(b)** Import flows.

The transshipment yard can be viewed as a queue system that performs several types of transshipment operations. Assuming that the system operates under some non-preemptive work conserving rule (see Federgruen & Groenevelt, 1986), the stability condition of the system implies that the total traffic intensity must be less than the number of servers. For an intermodal yard, this condition is stated as

$$(r_h^{\mathrm{i}}\lambda_h^{\mathrm{i}} + r_h^{\mathrm{e}}\lambda_h^{\mathrm{e}})T_q^{\mathrm{dt}} + (1-r_h^{\mathrm{i}})\lambda_h^{\mathrm{i}}(T_q^{\mathrm{ws}} + T_q^{\mathrm{st}}) + (1-r_h^{\mathrm{e}})\lambda_h^{\mathrm{e}}(T_q^{\mathrm{ts}} + T_q^{\mathrm{sw}}) < L_{hq}^1. \tag{19}$$

The first term in inequality (19) is the traffic intensity (i.e., the flow multiplied by the average transshipment or cycle time) of direct transshipment operations. This considers the expected processing time of a direct transshipment operation T_q^{dt}. The second term is the traffic intensity of indirect import operations, where T_q^{ws} and T_q^{st} are the expected cycle times of a wagon-to-storage and a storage-to-truck operation, respectively. The third term is the traffic intensity of indirect

export operations, where T_q^{ts} and T_q^{sw} are the expected processing times of a truck-to-storage and a storage-to-wagon operation, respectively. Finally, L_{hq}^1 is the number of rail-mounted gantry cranes for yard configuration $q \in Q$ and hub $h \in N$.

Inequality (19) is simplified to $a_{hq}^1 \lambda_h^{\text{e}} + b_{hq}^1 \lambda_h^{\text{i}} < L_q^1$, where

$$a_{hq}^1 = r_h^{\text{e}} T_q^{\text{dt}} + (1 - r_h^{\text{e}})(T_q^{\text{ts}} + T_q^{\text{sw}}), \tag{20a}$$

$$b_{hq}^1 = r_h^{\text{i}} T_q^{\text{dt}} + (1 - r_h^{\text{i}})(T_q^{\text{ws}} + T_q^{\text{st}}). \tag{20b}$$

In previous equations, a_{hq}^1 can be interpreted as the average unit processing capacity consumption of export operations. For its part, b_{hq}^1 is interpreted as the average unit processing capacity consumption of import operations. This is consistent with our previous definitions of capacity related parameters in Section 3. Given that λ_h^{e} and λ_h^{i} are related to the decision variables as in equations (2) and (3), and assuming node h is selected to be a hub, the form of the stability constraints in the hub location model under the hub configuration $q \in Q$ is

$$a_{hq}^1 \sum_{k \in K} \sum_{\substack{a \in A_k, \\ h = a_1}} w_k x_{ka} + b_{hq}^1 \sum_{k \in K} \sum_{\substack{a \in A_k, \\ h = a_2}} w_k x_{ka} \leq L_{hq}^1, \qquad \forall h \in N, q \in Q. \tag{21}$$

Inequalities (21) can also be referred to as processing rate capacity constraints. It is important to recognize that by the existence of indirect operations, requiring intermediate storage, the unit capacity consumptions of gantry cranes are increased.

Given that storage space is a limited resource and that storage operations have a notorious impact on the efficiency of rail mounted gantry cranes (Basallo-Triana et al., 2022), it is necessary to consider the interactions and capacity limitations regarding this area in the rail-road hub network design model. The requirements of storage space of export and import flows is determined according to the ITU dwell times, as we describe below.

Figure 4 shows a schematic representation of the yard accumulation and dissipation patterns for export and import ITUs. The batch of trains is available in the transshipment area at time t, and it takes on average a time T_{hq} for serving the batch, when the yard configuration is $q \in Q$ and for hub node $h \in N$. During the batch service, only a fraction of ITUs is transshipped directly; the remaining units are stored temporarily.

Let τ_{hq}^{e} and τ_{hq}^{i} be the average dwell times for export and import ITUs, respectively. Dwell times are defined for each yard configuration $q \in Q$ and hub $h \in N$ to guarantee that the fractions of direct transshipments remain unchanged, as it is assumed. Given that the fractions of direct

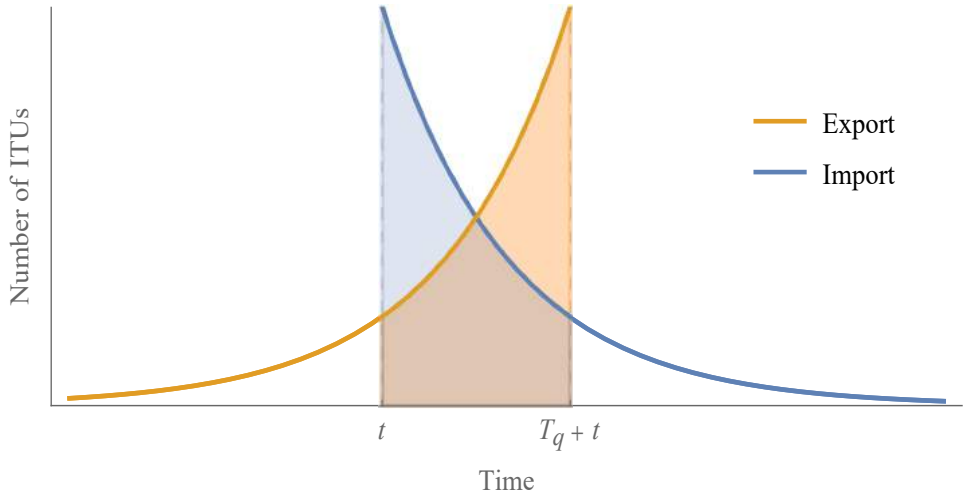

Figure 4: Accumulation of export ITUs and dissipation of import ITUs related to a batch of trains. Shaded areas are related to the fractions of direct transshipments.

transshipments are assumed to be known, we are interested in computing the associated ITU dwell times τ^{e}_{hq} and τ^{i}_{hq}.

Let us consider first the case of import operations. Import ITUs are available in the yard once the batch of trains is available for service. Without loss of generality, we consider that the batch service starts at time $t = 0$. Given that ITU dwell time is exponentially distributed, and assuming that transshipment times are negligible compared to the whole average batch service time T_{hq}, the fraction of direct transshipments r^{i}_h must satisfy the following relationship:

$$r^{\mathrm{i}}_h = \int_0^{T_{hq}} \frac{e^{-t/\tau^{\mathrm{i}}_{hq}}}{\tau^{\mathrm{i}}_{hq}} dt = 1 - e^{-T_{hq}/\tau^{\mathrm{i}}_{hq}}, \qquad \forall q \in Q, h \in Q. \tag{22}$$

Solving for τ^{i}_{hq} in equation (22) we obtain:

$$\tau^{\mathrm{i}}_{hq} = \frac{T_{hq}}{\ln\left(\frac{1}{1-r^{\mathrm{i}}_h}\right)}, \qquad \forall q \in Q, h \in Q. \tag{23}$$

Now consider the case of export operations. Export ITUs that arrive during the batch service time are directly transshiped to their respective train. By the uniform property of Poisson processes, truck arrival times follow a uniform distribution during the expected batch service time T_{hq} (Shortle et al., 2018, p. 43). Then, if the truck arrival time is x, the expected fraction of direct transshipments for export ITUs during the expected batch service time T_{hq} is given by

$$r^{\mathrm{e}}_h = \int_0^{T_{hq}} \frac{1}{T_{hq}} \left(\int_0^{T_{hq}-x} \frac{e^{-t/\tau^{\mathrm{e}}_{hq}}}{\tau^{\mathrm{e}}_{hq}} dt \right) dx = 1 - \frac{\tau^{\mathrm{e}}_{hq}}{T_{hq}} \left(1 - e^{-T_{hq}/\tau^{\mathrm{e}}_{hq}}\right), \qquad \forall q \in Q, h \in N. \tag{24}$$

Solving for τ^{e}_{hq} in equation (24) we obtain

$$\tau^{\mathrm{e}}_{hq} = \frac{(1 - r^{\mathrm{e}}_h)T_{hq}}{1 + (1 - r^{\mathrm{e}}_h)\mathcal{W}\left(\frac{e^{1/(r^{\mathrm{e}}_h - 1)}}{r^{\mathrm{e}}_h - 1}\right)}, \qquad \forall q \in Q, h \in N, \tag{25}$$

where $\mathcal{W}(\cdot)$ stands for the Lambert W function. It seems that we have ignored the stochastic nature of the batch service time when computing the average ITU dwell times in equations (23) and (25). However, as we show in Appendix A, the previous equations are still valid for stochastic batch service times.

Figure 5 shows the behavior of the fraction of direct transshipments as a function of the container dwell time for a given value of the batch service time T_{hq}. Import operations have a greater fraction of direct transshipments since import ITUs are available at the moment the batch is available for service. In contrast, the availability of export containers depends on the truck arrivals during the batch service time.

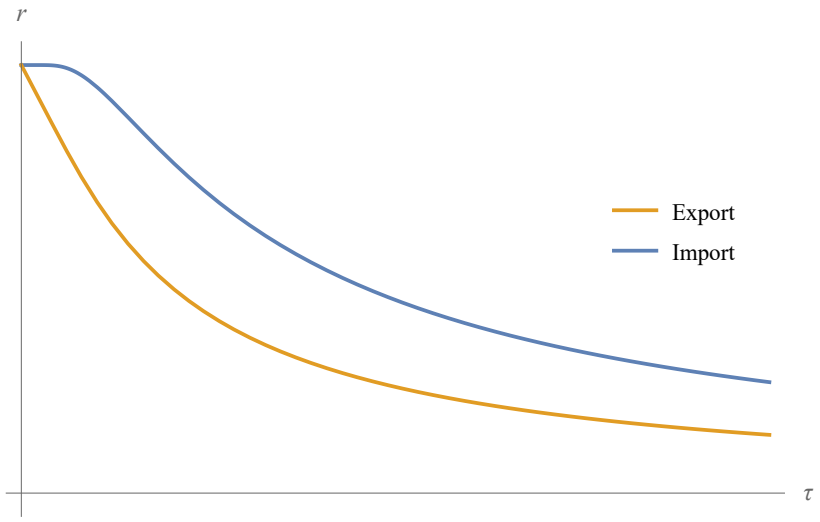

Figure 5: Fraction of direct transshipments as a function of the ITU dwell time.

The quantification of the expected batch service time T_{hq} depends on the average time for a direct transshipment operation T^{dt}_q, the average time for a wagon-to-storage operation T^{ws}_q, and the average time for a storage-to-wagon operation T^{sw}_q. These operations are carried out repeatedly according to the number of ITUs that need to be transshipped in the batch, which have an expected value of κ. The average batch service time also depends on the overall (weighted average) fraction of direct transshipments $\left(r^{\mathrm{e}}_h\lambda^{\mathrm{e}}_h + r^{\mathrm{i}}_h\lambda^{\mathrm{i}}_h\right) / \left(\lambda^{\mathrm{e}}_h + \lambda^{\mathrm{i}}_h\right)$, and on the number of rail mounted gantry cranes L^1_{hq}. According to this, the average batch service time is computed as

$$T_{hq} = \left[\frac{r^{\mathrm{e}}_h\lambda^{\mathrm{e}}_h + r^{\mathrm{i}}_h\lambda^{\mathrm{i}}_h}{\lambda^{\mathrm{e}}_h + \lambda^{\mathrm{i}}_h}T^{\mathrm{dt}}_q + \left(1 - \frac{r^{\mathrm{e}}_h\lambda^{\mathrm{e}}_h + r^{\mathrm{i}}_h\lambda^{\mathrm{i}}_h}{\lambda^{\mathrm{e}}_h + \lambda^{\mathrm{i}}_h}\right)\left(\frac{\lambda^{\mathrm{e}}_h}{\lambda^{\mathrm{e}}_h + \lambda^{\mathrm{i}}_h}T^{\mathrm{sw}}_q + \frac{\lambda^{\mathrm{i}}_h}{\lambda^{\mathrm{e}}_h + \lambda^{\mathrm{i}}_h}T^{\mathrm{ws}}_q\right)\right]\frac{\kappa}{L^1_{hq}}. \tag{26}$$

Equation (26) involves a complex non-linear function of the export and import flows to be considered in a hub location model. However, it is possible to simplify this expression. First, note that the expected cycle time of a storage-to-wagon operation is the same as the expected cycle time of a wagon-to-storage operation, i.e., $T_q^{\mathrm{sw}} = T_q^{\mathrm{ws}}$, since no rehandlings are expected to occur in export operations. Second, the overall fraction of direct transshipments $\left(r_h^{\mathrm{e}}\lambda_h^{\mathrm{e}} + r_h^{\mathrm{i}}\lambda_h^{\mathrm{i}}\right) / \left(\lambda_h^{\mathrm{e}} + \lambda_h^{\mathrm{i}}\right)$ can be simplified, with a loss of accuracy, to $(r_h^{\mathrm{e}} + r_h^{\mathrm{i}})/2$, in which we are assuming that $\lambda_h^{\mathrm{e}} = \lambda_h^{\mathrm{i}}$, which might not be the case in practice. We note that this loss of accuracy allows a great simplification of the hub location model. According to this, the expected batch service time T_{hq} is computed as

$$T_{hq} = \left[\frac{r_h^{\mathrm{e}} + r_h^{\mathrm{i}}}{2} T_q^{\mathrm{dt}} + \left(1 - \frac{r_h^{\mathrm{e}} + r_h^{\mathrm{i}}}{2}\right) T_q^{\mathrm{ws}}\right] \frac{\kappa}{L_{hq}^1}. \tag{27}$$

As before, it seems that we have ignored the stochastic nature of the time for a direct transshipment operation T_q^{dt}, the time for a wagon-to-storage operation T_q^{ws}, and the number of ITUs that need to be transshipped in a batch of trains κ. However, assuming independence of random variables, which is a reasonable assumption, equations (26) and (27) are valid, in probabilistic terms.

Equation (27) implies a loss of accuracy by assuming that $\lambda_h^{\mathrm{e}} = \lambda_h^{\mathrm{i}}$. To quantify the magnitude of the error, we note that when $\lambda_h^{\mathrm{e}} = 0$ the batch service time is $T_{hq} = r_h^{\mathrm{i}} T^{\mathrm{dt}} + (1 - r_h^{\mathrm{i}}) T^{\mathrm{ws}}$. For its part, when $\lambda_h^{\mathrm{e}} = 0$ the batch service time is $T_{hq} = r_h^{\mathrm{e}} T^{\mathrm{dt}} + (1 - r_h^{\mathrm{e}}) T^{\mathrm{ws}}$. The absolute value of the difference between previous quantities is $(T_q^{\mathrm{ws}} - T_q^{\mathrm{dt}})|r_h^{\mathrm{e}} - r_h^{\mathrm{i}}|$. Then, the maximum associated error of can be computed as

$$\frac{(T_q^{\mathrm{ws}} - T_q^{\mathrm{dt}})|r_h^{\mathrm{e}} - r_h^{\mathrm{i}}|}{2}. \tag{28}$$

Consider, for example, a yard operating 2 cranes and a storage block of width 4 ITUs and a maximum stacking height of 4 ITUs. Assume an extreme situation in which $|r_h^{\mathrm{e}} - r_h^{\mathrm{i}}| = 1$ and $\lambda_h^{\mathrm{e}} = 0$ or $\lambda_h^{\mathrm{i}} = 0$. Considering data from Basallo-Triana et al. (2022), the maximum error is 0.51 minutes per movement of a gantry crane. This error decreases as the number of cranes increases. Of course, a more realistic scenario may consider $|r_h^{\mathrm{e}} - r_h^{\mathrm{i}}| \approx 0$ and $\lambda_h^{\mathrm{e}} \approx \lambda_h^{\mathrm{i}}$, implying a smaller error.

To quantify the use of the storage space, it is necessary to determine the average time the ITUs spend on the storage yard. Consider the case of import ITUs. Recall that our definition of dwell time includes the time an ITU spends in a rail wagon and, if necessary, the time the ITU spends in the storage yard. Then, the expected value of the time the ITU spends in the storage yard is computed using the conditional expectation $E[t|t \geq T_{hq}] - T_{hq}$, where t is the random variable for the container dwell time. By the memoryless property of the exponential distribution, we have

$E[t|t \geq T_{hq}] - T_{hq} = \tau^{\mathrm{i}}_{hq}$, i.e., the average storage time coincides with the average dwell time. The same situation occurs for export operations.

Now we are in a position to establish storage space capacity constraints. The ITU arrival rate to the storage yard of hub $h \in N$ is $(1 - r^{\mathrm{i}}_h)\lambda^{\mathrm{i}}_h + (1 - r^{\mathrm{e}}_h)\lambda^{\mathrm{e}}_h$. By Little's law, the average ITU inventory level in the storage yard is $(1 - r^{\mathrm{i}}_h)\lambda^{\mathrm{i}}_h\tau^{\mathrm{i}}_{hq} + (1 - r^{\mathrm{e}}_h)\lambda^{\mathrm{e}}_h\tau^{\mathrm{e}}_{hq}$. If the maximum expected storage space utilization in yard configuration $q \in Q$ is $u_{hq} \in [0, 1]$, and the total storage space available is E_{hq}, then the maximum effective storage space is $L^2_{qh} = u_{hq}E_{hq}$. Thus storage space capacity constraints under hub configuration q can be defined as:

$$a^2_{hq} \sum_{k \in K} \sum_{\substack{a \in A_k, \\ h = a_1}} w_k x_{ka} + b^2_{hq} \sum_{k \in K} \sum_{\substack{a \in A_k, \\ h = a_2}} w_k x_{ka} \leq L^2_{hq}, \qquad \forall h \in N, q \in Q, \tag{29}$$

where

$$a^2_{hq} = (1 - r^{\mathrm{e}}_h)\tau^{\mathrm{e}}_{hq}, \tag{30a}$$

$$b^2_{hq} = (1 - r^{\mathrm{i}}_h)\tau^{\mathrm{i}}_{hq}. \tag{30b}$$

Here, a^2_{hq} can be interpreted as the unitary consumption of storage space of export ITUs, and b^2_{hq} as the unitary consumption of storage space of import ITUs.

Despite the fact that generalized capacity constraints are obtained from a stochastic analysis of crane processing times, ITU dwell times and truck arrival patterns, such constraints are not stochastic in nature. The establishment of generalized capacity constraints, which are satisfied with some probability, requires the analysis of the probability distributions associated to the parameters.

5. Computational experiments

Our formulations were coded in C using the callable library of CPLEX 20.1.0 and executed on a PC with an Intel Core i7 processor running at at 2.90 GHz and 32.0 GB of RAM under a Windows 10 Pro, 64 bits, environment. We set a time limit for the solution to 86,400 seconds.

We use about 380 problem instances from the well-known AP data set with up to 60 nodes. We also introduce the COL data set, a new data set for intermodal hub location considering freight transportation data from Colombia. We use about 300 problem instances of the COL data with up to 60 nodes. Our interest is the analysis of the performance of the `BM` and `MC` formulations. We also analyze the impact of the level of synchronization at hub facilities. Finally, we discuss relevant insights from the Colombia case study for intermodal transportation.

5.1. AP data set

The first set of computational experiments was performed on the well known Australian Post (AP) data (Ernst & Krishnamoorthy, 1996). The data set provides the installed capacity for each potential hub location, which is denoted by μ_h, for $h \in N$. The data set also provides the fixed installation costs, denoted by f_h, for $h \in N$. Capacities and fixed costs are defined in two different scenarios, which are referred to as loose (L) and tight (T), respectively. We start by defining capacity-related parameters as follows:

$$a_h^r = \frac{1}{\mu_h^{l^r}}, \qquad b_h^r = \frac{1}{0.9\mu_h^{l^r}}, \qquad \text{for } r = 1, h \in N, \tag{31}$$

$$L_h^r = \mu_h^{(1-l^r)}, \qquad \text{for } r = 1, \ldots, |R|, h \in N, \tag{32}$$

where l^r is a uniformly distributed random number with density function $l^r \sim U(0,1)$, for $r = 1, \ldots, |R|$. The generation of capacity consumption parameters for additional resources is performed as follows:

$$a_h^r = \frac{\sqrt{2} l_h^r L_h^r}{\mu_h}, \qquad b_h^r = \frac{\sqrt{2(1-(l_h^r)^2)} L_h^r}{0.9\mu_h}, \qquad \forall h \in N, r = 2, \ldots, |R|, \tag{33}$$

where l_h^r is a uniformly distributed random number with distribution $U(0,1)$. As a consequence of the previous definitions of capacity-related parameters, the feasible region generated by generalized capacity constraints defines a (convex) outer approximation of the elliptic region $(\lambda_h^{\mathrm{e}})^2 + (\lambda_h^{\mathrm{i}}/0.9)^2 \leq \mu_h^2/2$. In this case, all resources in R are dominant. It is worth noting that the generalized capacity constraint associated with each additional resource reduces the size of the feasible region for the export and import flows passing through a hub.

We generate capacity related parameters for different hub configurations as follows:

$$a_{h|Q|}^r = a_h^r, \qquad b_{h|Q|}^r = b_h^r, \qquad L_{h|Q|}^r = L_h^r, \qquad \forall r \in R, h \in N, \tag{34}$$

$$a_{hq}^r = \frac{a_{h,q+1}^r}{0.7^{l_r}}, \qquad b_{hq}^r = \frac{b_{h,q+1}^r}{0.7^{l_r}}, \qquad L_{hq}^r = 0.7^{(1-l_r)} L_{h,q+1}^r, \qquad \forall r \in R, h \in N, q = 1, \ldots, |Q|-1. \tag{35}$$

Finally, the fixed installation costs for the different hub configurations are obtained as $f_{h|Q|} = f_h$ and $f_{hq} = 0.9 f_{h,q+1}, h \in N, q = 1, \ldots, |Q|-1$.

For the COL data set, we set the unit transport costs as follows $\chi = 3, \tau = 0.75, \delta = 2, \rho = 7$. Also, the hub capacities μ_h from the original data set are multiplied by a factor of 0.5.

All problem instances were solved to optimality within a time limit of 86,400 seconds. The maximum time required to solve an instance of the `BM` formulation is 44,389 seconds, which corresponds to the instance 60LT with $|R| = 2$ resources and $|Q| = 2$ hub configurations. For its part,

the maximum time required to solve an instance of the MC formulation is 70,897 seconds, which corresponds to instance 60LT with $|R| = 3$ resources and $|Q| = 2$ hub configurations. The optimal number of hubs varies from 2 to 8. More hubs seem to be required for the LL and LT instances, while fewer hubs are required for the TL and TT instances. The number of hubs tend to decrease when the number of nodes in the network increases.

Figure 6 shows a paired comparison of the total solution time of BM and MC formulations. When $|N| = 10$, the MC formulation solves the problem more efficiently in 73 % of the instances. This quantity becomes 27 %, 70 %, 51 %, 60 % and 45 % for $|N| = 20, |N| = 25, |N| = 40, |N| = 50$ and $|N| = 60$, respectively. In general, the MC formulation solves more efficiently a greater number of instances. According to a Wilcoxon signed rank test, the MC formulation outperforms the BM formulation in terms of solution time for instances with 10 and 25 nodes. For its part, the BM formulation outperforms the MC formulation in the 20-node instances. Previous comparisons reach p values of the order of 10^{-4}. The instances with 40, 50 and 60 nodes do not produce significant differences in the processing times of the formulations. Despite the MC formulation having theoretical advantages in terms of formulation strength, it seems that the increased size of the formulation may affect its performance in the context of the AP data set.

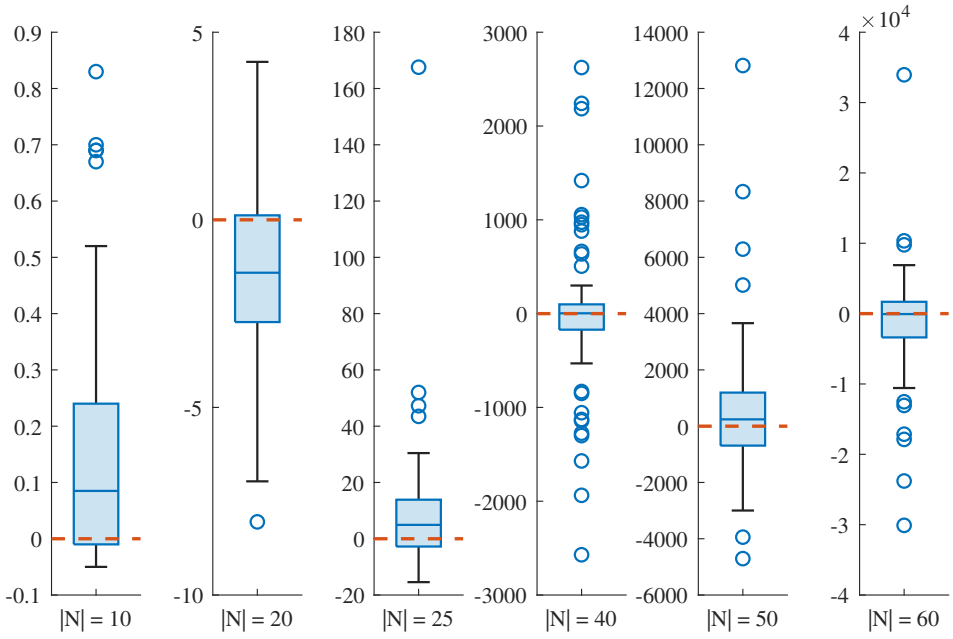

Figure 6: Distribution of the difference between the total solution time of the BM formulation and the total solution time of the MC formulation for the AP data set. Differences are given in seconds.

The previous findings can also be observed in Figure 7. We note that when $|Q| = 1$, both formulations `BM` and `MC` perform very similarly in most instances. This is reasonable since when $|Q| = 1$ both formulations are essentially the same in their basic structure, with the difference that the `MC` formulation has redundant auxiliary continuous variables. When $|Q|$ increases, the differences between the methods become more obvious. When $|R|$ increases, average processing times between the methods are similar. It can be seen that increasing $|Q|$ has a greater impact on the average solution time than does increasing $|R|$.

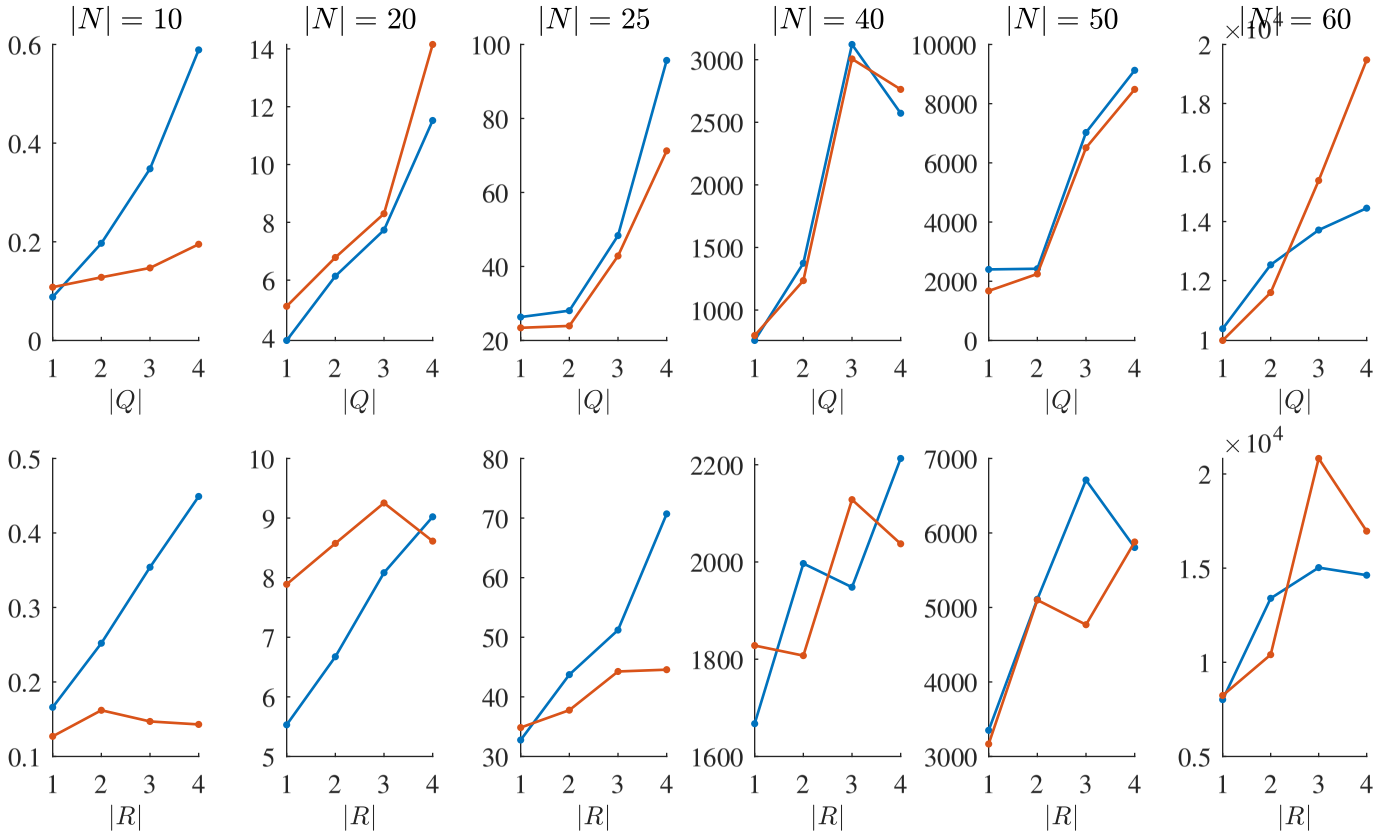

Figure 7: Average solution time in seconds for different values of $|Q|$ and $|R|$. **Blue**: `BM` formulation. **Orange**: `MC` formulation.

The largest linear programming (LP) relaxation gap obtained in the context of the AP data set is 2.07% and was reached for the 40LL instance with $|R| = 1$ and $|Q| = 3$, for both `BM` and `MC` formulations. In general, we get a small LP relaxation gap, which tends to decrease with the network size. Interestingly, both the `BM` and `MC` formulations have the same LP relaxation gap. This fact is surprising since one can expect a greater LP relaxation gap for the `BM` formulation when compared to the `MC` formulation because the `MC` formulation is stronger than the `BM` formulation in the sense of the tightness of generalized capacity constraints. However, it is necessary to recall that this is just from the point of view of the generalized capacity constraints. The LP relaxation gap does not depend exclusively on the generalized capacity constraints. The remaining constraints cause a greater contribution to the gap.

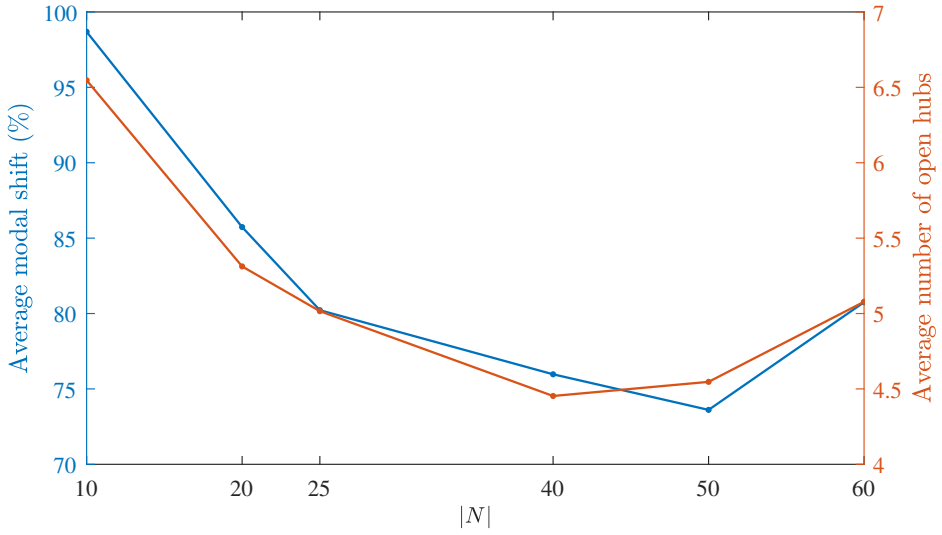

Figure 8: **Left axis**: Average percentage of flow captured by the intermodal transportation system in the AP data set. **Right axis**: Average number of open hubs.

Figure 8 shows the impact of the intermodal transportation system on the transportation of flow between origin and destination nodes. The average percentage of flow captured by the intermodal transportation system, also referenced to as modal shift percentage, is high for small network sizes. This is because a small network contains nodes distant from each other, implying expensive direct transport operations compared to intermodal transport operations. When the number of nodes increases, then the distances between a node and other nodes in its neighborhood decrease, and direct transport operations between those nodes become more attractive. Interestingly, the modal shift percentage decreases up to 50 nodes and then increases. We also note that the number of open

hubs has a high correlation with the modal shift percentage, as it is expected in order to capture more flow.

To obtain accurate intermodal hub network designs one needs a network with a sufficiently large size. This allows to determine a correct modal shift percentage and a suitable number of hubs. This is of interest in practical applications, for example, from an environmental point of view (Basallo-Triana et al., 2021; Bouchery & Fransoo, 2015). In this sense, unit transport costs χ, τ, δ and ρ should consider not only the internal costs of each mode of transport, but also the external cost associated to environmental and social impacts. It is recognized that a very high modal shift, as it occurs when direct transportation decisions are disregarded, may bring negative consequences in terms of costs and environmental impact (Bouchery & Fransoo, 2015; Taherkhani & Alumur, 2019).

5.2. Colombian case study

We introduce the *COL data set*, a new data set for intermodal hub location that considers real freight transportation data from Colombia. The data was obtained from the National Freight Transport Registry (RNDC, by its acronym in Spanish), which is administered by the Colombian Ministry of Transport (Mintransporte). We collected data for the year 2019, which accounts for a total transport flow of more than 118 million tons. The current intermodal transport system in Colombia is precarious, and only about 1% of the freight uses the intermodal transport option. For this reason, and anticipating a more intense intermodal scenario, the demand w_k for commodity k represents the 30 % of the actual transport demand of such commodity. Then, we express flow data in ITUs per year, considering an average ITU weight of 12.3 tons.

The flow matrix comprises the flows between 1096 nodes, which correspond to different municipalities around the country. This is a sparse matrix in which more than 95% of the OD flows are zero. To produce instances with a smaller number of nodes, we aggregate geographical data using a partitioning around medoids (PAM) clustering algorithm (Kaufman & Rousseeuw, 2009, p. 68–104). The algorithm considers the haversine distance between nodes given their coordinates in latitude and longitude. In the implementation of the algorithm, we also consider the relevance of each node (municipality) in terms of the total inbound and outbound flow. In this sense, nodes with high total inbound and outbound flow have higher possibilities to be cluster representatives. Each smaller instance contains only the flows between cluster representatives, which are aggregated from the flows of the municipalities that belong to that cluster. The distribution of the nodes in

the COL data set is shown in Figure 9 along with the 60 representative municipalities.

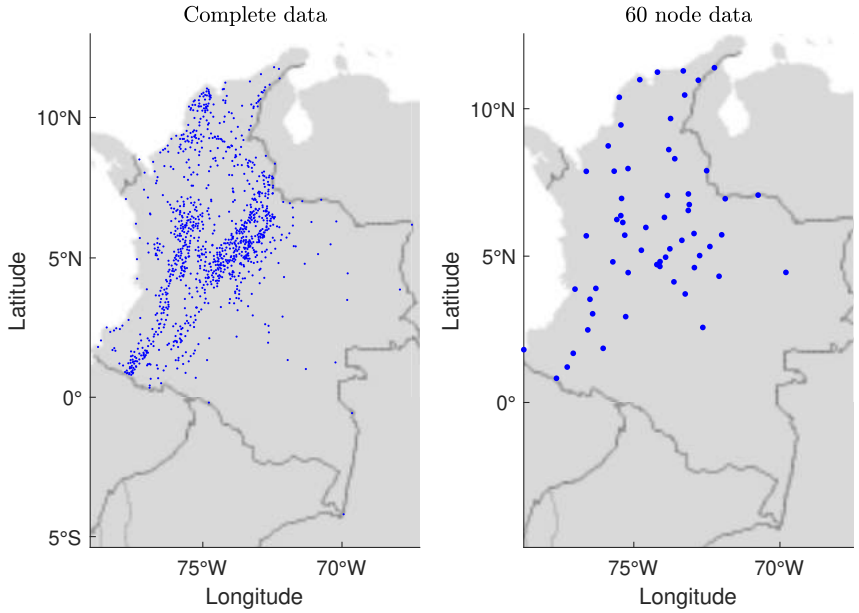

Figure 9: Colombia data set for intermodal hub location.

The hub facility is a rail-road transshipment yard (see Section 4.1), which is characterized by the following:

- The yard length is fixed to 600 m,
- The number of rail-mounted gantry cranes varies depending on the scenario considered,
- The maximum stacking height varies depending of the scenario considered,
- The storage width varies depending of the scenario considered,
- The storage space utilization is set to $u_{hq} = 0.6$,
- The fractions of direct transshipments r_h^{e} and r_h^{i}, $h \in N$ vary depending on the scenario considered.

In the data set, potential installed capacities are assumed to be the same for all nodes, i.e., $L_{hq}^r = L_q^r$, for $h \in N, q \in Q, r \in R_q$. The annual cost of investment in facilities is estimated using data from (Bontekoning, 2006, p. 112–116), but it is corrected by a factor that takes into consideration the variations in the construction costs between different nodes.

As we discuss in Section 4, for a rail-road transshipment yard, there are $|R| = 2$ resources of interest including the cranes and the storage space availability. To understand the dominance

relations between these resources, we conduct an analysis by varying the number of cranes from 1 to 4, the maximum stacking height from 1 to 4, the storage width from 2 to 4, and the the fractions of direct transshipments are defined as $r^{\mathrm{e}}, r^{\mathrm{i}} \in \{0.2, 0.3, \ldots, 0.9, 1.0\}$. According to this, we obtain a total of 3,888 different yard configurations. In 95.6 % of the configurations the cranes are the only dominant resource. In 4.2 % of the configurations both the cranes and the storage space availability are dominant. Finally, in 0.2 % of the configurations the storage space availability is the only dominant resource.

For the COL data set, we set the unit transport costs as follows $\chi = 1, \tau = 0.75, \delta = 1, \rho = 1.5$.

Scenario 1

The purpose of this scenario is to evaluate the performance of the `BM` and `MC` formulations in a situation of abundant storage space availability. In this sense, the cranes are the only dominant resource, i.e, $|R_q| = 1$, and storage space constraints are not required, which is the most likely scenario, as we discussed before. However, it is necessary to recall that the presence of indirect operations due to the storage of ITUs reduces the performance of the cranes. Yard configurations consider that the number of cranes varies from $|Q| = 1$ to $|Q| = 4$ while the maximum stacking height and the width of the storage area are fixed to 4 ITUs. We assume a constant storage space utilization of 0.6 to quantify the processing time of cranes.

Parameters a^r_{hq} and b^r_{hq} are computed using equations (20) and (30), for $r = 1, 2$. The expected cycle times of gantry cranes are obtained according to Basallo-Triana et al. (2022). For their part, the fractions of direct transshipments r^{e}_h and r^{i}_h are randomly generated from the uniform distribution $U(0.2, 1.0)$, for each node $h \in N$. We tested problem instances with 10, 20, 25, 40, 50 and 60 nodes. Each network size is evaluated at 25 random realizations of the fractions of direct transshipments except from the network with 60 nodes where only 19 problem instances where considered, implying a total of 144 problem instances.

The `BM` could not solve five problem instances within a time limit of 86,400 seconds. These instances are in the 60 node set reaching a maximum optimality gap of 3.60 %. For its part, the `MC` could not solve four problem instances in the 60 node set reaching a maximum optimality gap of 1.67 %. The optimal number of hubs varies from 6 to 8, and in most of the cases 6 hubs are open. The number of hubs tend to decrease as the network size increase.

Figure 6 shows a paired comparison of the total solution time of `BM` and `MC` formulations. Except for the 20-node instances, there is significant statistical evidence that suggest that the `MC`

outperforms the BM formulation in terms of processing time, with a p value of at most 0.02. When $|N| = 10$, the MC formulation solves the problem more efficiently in 100 % of the instances. This quantity becomes 60 %, 84 %, 76 %, 88 % and 79 % for $|N| = 20, |N| = 25, |N| = 40, |N| = 50$ and $|N| = 60$, respectively. This data set exploits in a better way the advantages of the MC formulation in terms of formulation strength.

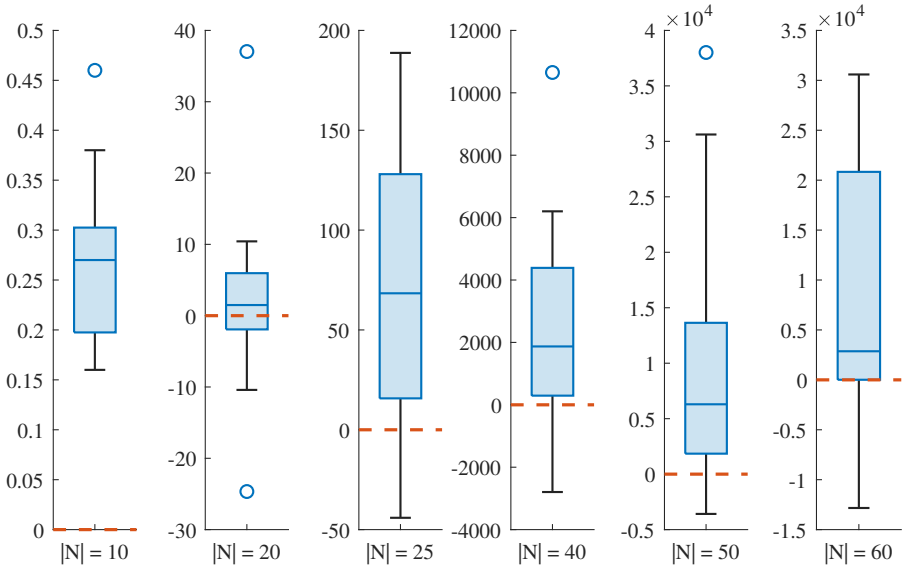

Figure 10: Distribution of the difference between the total solution time of the BM formulation and the total solution time of the MC formulation for the COL data set. Differences are in seconds.

The maximum LP relaxation gap is 2.82%, reached for instance 20-24. In general, the LP relaxation gap is the same for the BM and MC formulations, as we discussed for the AP data set. The LP relaxation gap is uniform as the network size increases.

Figure 11 shows the impact of the intermodal transportation system for the COL data set. As in the case of the AP data set (see Figure 8), it is observed that the modal shift decreases as the number of nodes in the network increases because the closeness between neighboring nodes increases. As before, there is a high positive correlation between the number of open hubs and the modal shift. There is a slight increase in the average modal shift when the network size changes from 50 to 60 nodes.

Scenario 2

This scenario aims to evaluate the impact of the level of synchronization on total transportation and investment costs. We consider problem instances for different combinations of the fractions

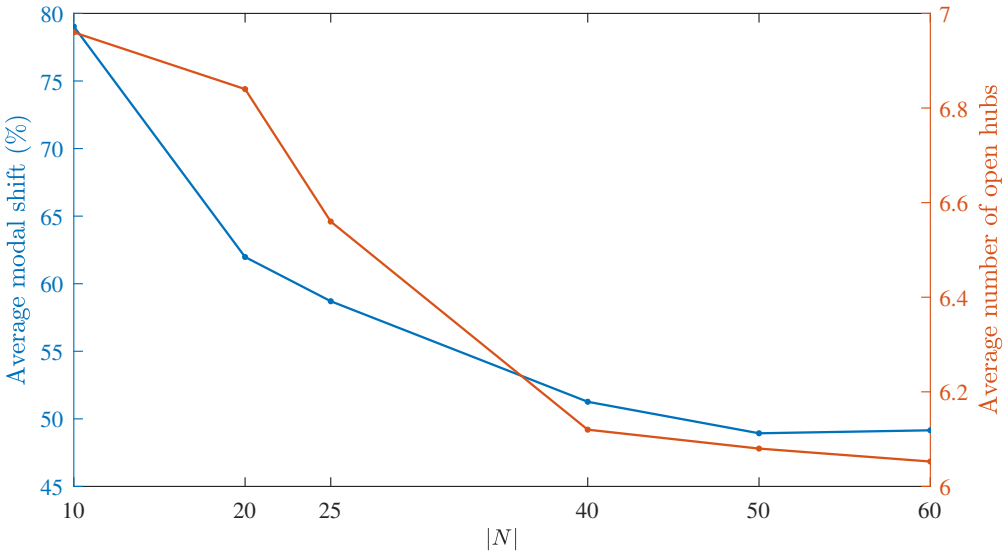

Figure 11: **Left axis**: Average percentage of flow captured by the intermodal transportation system in the COL data set. **Right axis**: Average number of open hubs.

of direct transshipments as follows: $r_h^{\mathrm{e}} = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $r_h^{\mathrm{i}} = \{0.2, 0.4, 0.6, 0.8, 1.0\}$. This represents a total of 25 possible combinations. In this context, all nodes $h \in N$ share the same fractions of direct transshipments corresponding to the specific combination of r^{e} and r^{i}. We evaluate each combination on problem instances with 10, 20, 25 and 40 nodes, and report average costs.

The level of synchronization has a significant impact on total costs (see Figure 12 (a)). The less synchronized the operations, the higher the total investment cost on gantry cranes. In a situation with very low synchronization, i.e., $r^{\mathrm{e}} = r^{\mathrm{i}} = 0.2$, the average number of cranes installed at each hub is 2.68. In contrast, in a situation of high synchronization ($r^{\mathrm{e}} = r^{\mathrm{i}} = 1.0$), the average number of cranes installed is 1.29 (see Figure 12 (b)). This is explained since low values for r^{i} and r^{e} lead to an increase in the interactions with the storage area of the yard reducing the processing capacity of cranes. Then, more resources (cranes) need to be added to satisfy the transshipment demand. It can be observed that total cost increases more rapidly when r^{i} decrease than when r^{e} decrease. This is because import operations consume more resources than export operations as a consequence of rehandlings.

The cost structure and modal shift for different values of the fractions of direct transshipments are shown in Figure 13. As the fractions of direct transshipment increase, the investment in hub facilities and the direct transport cost decrease. For its part, the intermodal transport cost along

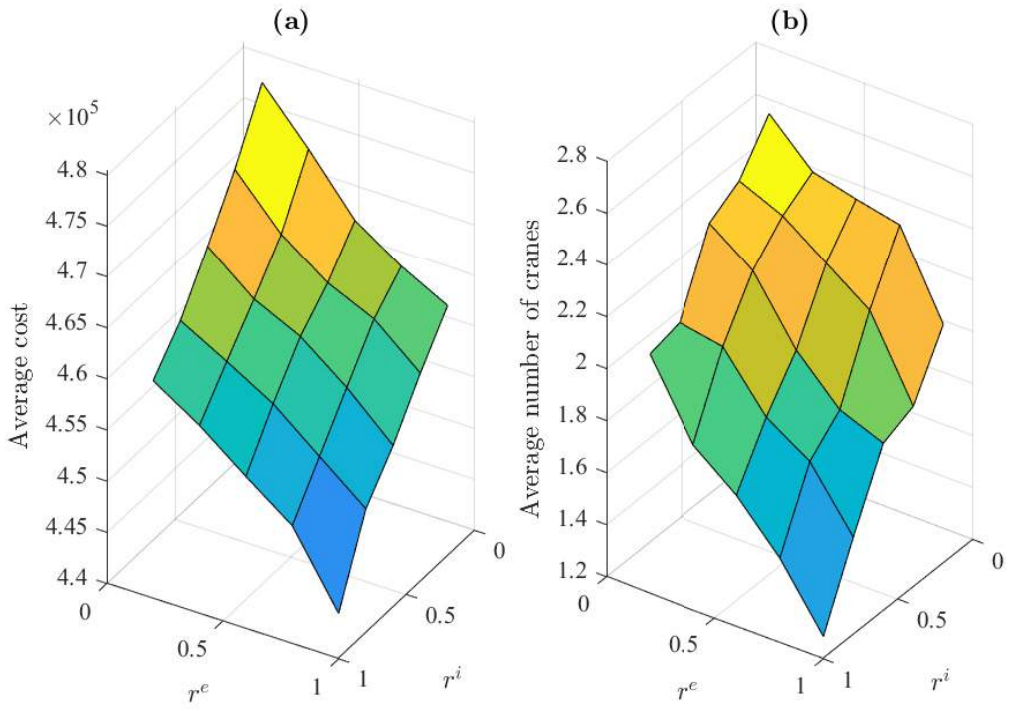

Figure 12: **(a)** Average costs for different combinations of the fractions of direct transshipments. **(b)** Average number of cranes per hub.

with the modal shift is increased as the fraction of direct transshipment decrease. There is a high positive correlation between the intermodal transport cost and modal shift. There is a negative correlation of previous variables with direct transport costs. Clearly, the intermodal transport is enhanced when highly efficient transshipment yards are operating.

Scenario 3

The purpose of this scenario is to evaluate the performance of the formulations in cases of limited storage space availability. In these situations, storage space capacity constraints become dominant along with processing capacity constraints. Then, the number of resources to be considered is $|R_q| = 2$, for all $q \in Q$, and the number of hub configurations varies from $|Q| = 8$ to $|Q| = 16$. Hub configurations refer to different yard designs by varying aspects such as the number of cranes, the maximum stacking height, and the width of the storage area. We evaluated 31 problem instances for network sizes of up to 40 nodes. Most of the instances with 40 nodes cannot be solved within 84,600 seconds, and we show the results for some of such instances.

The results of this scenario are shown in Appendix B. It is worth noting that this scenario

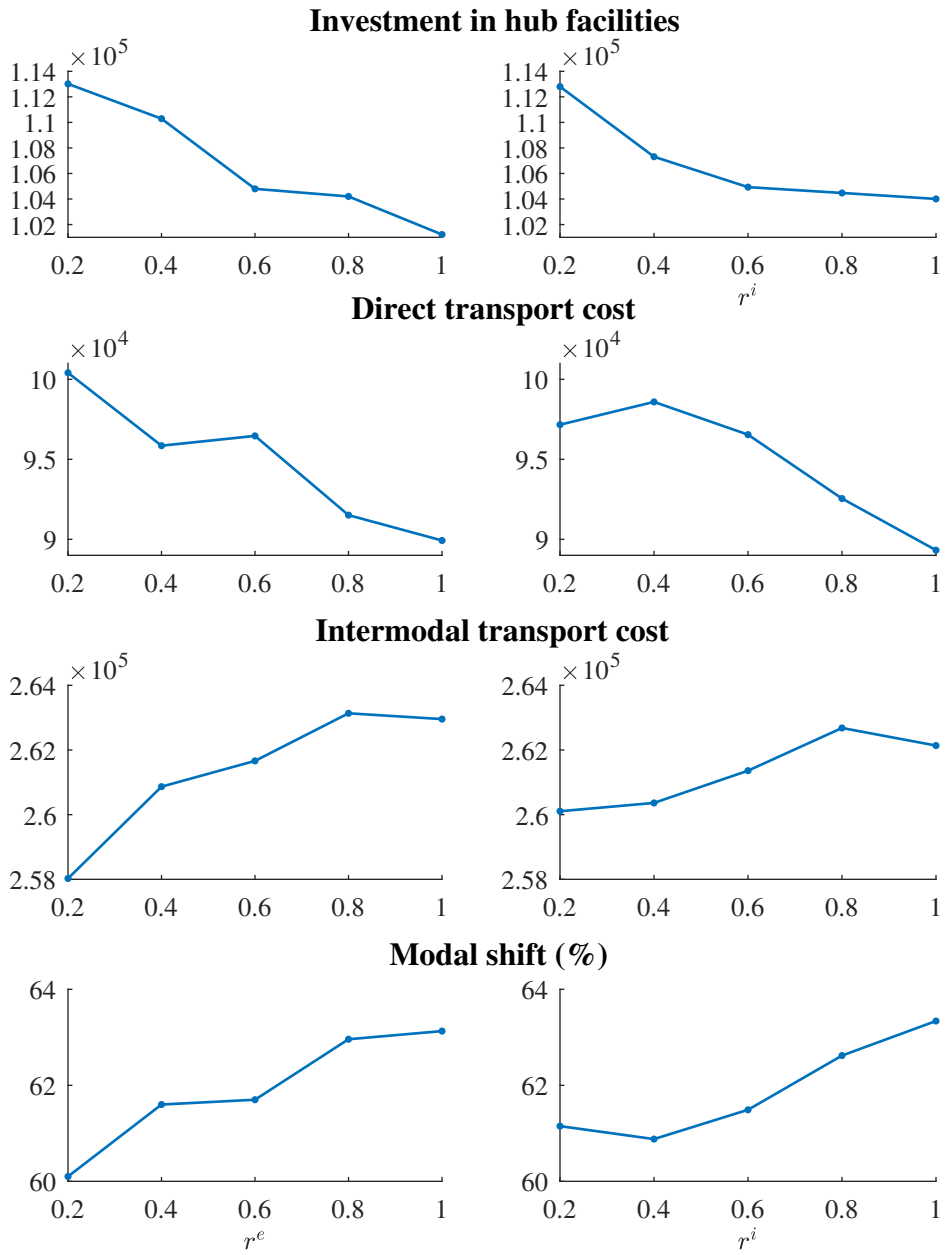

Figure 13: Variations on costs and modal shift for different values of the fraction of direct transshipment of export operations (left panels) and import operations (right panels). Values are averaged over the 10, 20, 25 and 40 node instances.

contains much more challenging instances than those of previous scenarios. The MC formulation outperforms BM formulation in 29 out of 31 problem instances in terms of processing time. There is a clear superiority of the MC formulation, which is explained by the increased formulation strength. Interestingly, in the majority of the cases, the BM explores fewer nodes in the search tree of the

branch and bound algorithm than the MC formulation, this behaviour is also present in previous scenarios and the corresponding results of the AP data set. Regarding the LP relaxation gap, we note that there is no difference between the gaps of the BM and MC formulations as discussed before.

Managerial insights for the Colombian case study

We concentrate on the 60-node instances. Figure 14 shows the most frequent hub configuration. The selected municipalities are located in important regions for logistics operations in the country. Guadalajara de Buga (G. Buga) is located in Valle del Cauca, which is relevant because it contains one of the most important ports of the country, the port of Buenaventura, performing logistics operations through the Colombian Pacific Region. Guadalajara de Buga is the most prominent municipality for the location of a rail-road transshipment yard because of their more strategic position to connect the center of the country with the Pacific Region. This is in agreement to the National Logistics Policy (CONPES 3547 document) from the Colombian Government.

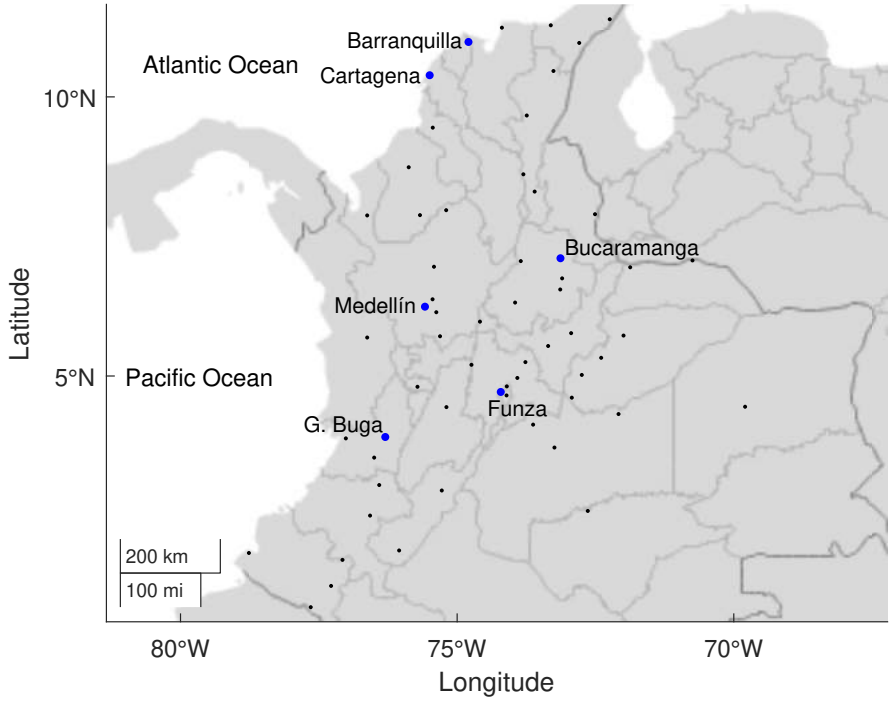

Figure 14: Most frequent hub configuration in the 60-node data set for scenario 1.

Cundinamarca and Bogotá constitute an important central region for logistics operations in the country. Most of the transport flows in the country are originated or destined to such departments. Funza is located in Cundinamarca and has great proximity to Bogotá, This municipality is frequently selected to be a hub. Funza constitute a potential location of a rail-road transshipment

facility by connecting the center of the country with the most important seaports in the Pacific Region. It also relevant to connect important municipalities such as Meddellín (Cundinamarca) and Bucaramanga (Santander).

The department of Bolivar contains the Port of Cartagena, the department of Atlántico contains the port of Barranquilla. These are important ports connecting Colombia to the international traffic in the Atlantic Ocean. Cartagena and Barranquilla are selected as rail-road transshipment facilities mainly because of the presence of important seaports in these municipalities. These municipalities serve as strategic position for connecting the center of the country to the ports in the Atlantic Region.

Generalized capacity constraints are of the form $a^1_{hq}\lambda^{\mathrm{e}}_h + b^1_{hq}\lambda^{\mathrm{i}}_h < L^1_{hq}$, where L^1_{hq} is the number of gantry cranes and a^1_{hq} and b^1_{hq} are the weighted average crane processing time for export and import operations, respectively, which are weighted according to the fractions of direct transshipment following the equations in (20). Parameters a^1_{hq} and b^1_{hq} are defined according to the expected crane processing time; however, we note that crane processing times are random in nature. In order to produce more robust hub network designs, we associate a beta probability distribution to the weighted processing times of import and export operations, respectively. The parameters of the beta distributions are obtained following Basallo-Triana et al. (2022). Then we consider the generalized capacity constraints $a^{1p}_{hq}\lambda^{\mathrm{e}}_h + b^{1p}_{hq}\lambda^{\mathrm{i}}_h < L^1_{hq}$, where a^{1p}_{hq} and b^{1p}_{hq} are the p-th percentile of the corresponding beta distribution.

Table 1 shows a sensitivity analysis by varying the unit capacity consumption percentiles, which are shown in the second column, and considering 40-node instances. The third column shows the average number of cranes of the installed hubs. The column under the heading *Average hub utilization (%)* shows the average hub utilization, which is computed using the expected value of the unit capacity consumption parameters as follows $(a^r_{hq}\lambda^{\mathrm{e}}_h + b^r_{hq}\lambda^{\mathrm{i}}_h)/L^r_{hq}$. The optimal objective value is reported in the column under the heading *Objective value*. The column under the heading *Modal shift (%)* corresponds to the percentage of flow that is captured by the intermodal transportation system. The optimal hubs and the corresponding hub configurations are shown in the last column of the table.

Increasing the unit capacity consumption percentiles has an analogous impact to reducing the capacity availability, implying tighter capacity constraints. In this scenario, the total cost increases mainly because of higher investment in yard cranes. Considering higher percentiles for unit capacity consumption parameters also reduces the expected hub utilization, as it is expected. This is a good

practice to prevent over-congested hubs and to obtain more robust hub configurations. We also note that varying capacity consumption percentiles does not notoriously affect the modal shift.

Table 1: Sensitivity analysis considering different percentiles of the unit capacity consumption parameters a^r_{hq} and b^r_{hq} for 40-node instances.

r^e, r^i	Capacity percentile	Average number of cranes	Average hub utilization (%)	Objective value	Modal shift (%)	Optimal hub - Hub configuration
0.2, 0.2	Median	2.5	100	496,008	47	1-2, 4-3, 19-3, 31-3, 32-2, 34-2
	Mean	2.7	97	497,290	50	1-2, 4-3, 19-4, 31-3, 32-2, 34-2
	60	2.7	94	499,572	48	1-2, 4-3, 19-4, 31-3, 32-2, 34-2
	70	3.0	87	501,684	51	1-3, 4-3, 19-4, 29-2, 31-4, 34-2
	80	3.0	82	504,678	49	1-3, 4-3, 19-4, 29-2, 31-4, 34-2
	90	3.2	75	509,511	47	1-3, 4-3, 19-4, 29-2, 31-4, 34-3
0.6, 0.6	Median	1.8	97	482,334	48	1-2, 4-2, 19-2, 29-1, 31-2, 34-2
	Mean	2.0	95	483,543	47	1-2, 4-2, 19-2, 29-1, 31-2, 34-2
	60	2.2	84	484,001	52	1-2, 4-2, 19-3, 29-1, 31-3, 34-2
	70	2.3	77	484,635	53	1-2, 4-2, 19-3, 31-3, 32-2, 34-2
	80	2.3	76	486,133	52	1-2, 4-2, 19-3, 31-3, 32-2, 34-2
	90	2.5	68	490,692	51	1-2, 4-3, 19-3, 31-3, 32-2, 34-2
1.0, 1.0	Median	1.2	94	464,443	55	1-1, 4-1, 19-2, 31-1, 32-1, 34-1, 38-1
	Mean	1.1	92	465,841	54	1-1, 4-1, 19-2, 31-1, 32-1, 34-1, 38-1
	60	1.3	80	468,218	55	1-1, 4-1, 19-2, 25-1, 31-2, 32-1, 34-1
	70	1.4	70	470,718	57	1-1, 4-2, 19-2, 31-1, 32-1, 34-1, 37-2
	80	1.7	58	472,692	52	1-2, 4-2, 19-2, 31-2, 32-1, 34-1
	90	1.7	48	474,427	52	1-2, 4-2, 19-2, 29-1, 31-2, 34-1

6. Conclusions

We have proposed a new hub location modeling approach by including generalized capacity constraints. This approach recognizes the existence of different types of limited hub resources and that import and export flows have different unit capacity consumptions. The previous features allow more realistic modeling of hub facilities. We present two path-based formulations for the hub location model with generalized capacity constraints called big-M (`BM`), which is a small formulation, and multiple choice (`MC`) formulation, which is a larger but stronger formulation. After evaluating the performance of the formulations in the AP, we found that the `MC` formulation tends to outperform the `BM` in terms of solution times.

We carried out an analysis of train-truck synchronization from a stochastic point of view to understand how synchronization may affect the hub network structure. The consideration of synchronization aspects allowed us to show the necessity of generalized capacity constraints in order to model more realistically a rail-road transshipment yard. It was shown that the interactions with a storage area, which is a limited resource, affect the yard performance, and that the processing rates of export and import flows are affected in different ways. It was observed that the level of synchronization has a notorious impact on the structure of the transshipment yard increasing investment costs, which are mainly affected by import operations. Synchronization affects intermodal and direct transport costs in different ways. The higher the synchronization level, the smaller the direct transport cost and the higher the intermodal transport cost.

For intermodal hub network design, it is recommended to have a network of a considerable size. This allows to determine a suitable modal shift and an appropriate number of hubs.

Our methodology is suitable for the modeling of more complex hub facilities. For example, seaports operate with many resources including quay cranes, yard cranes, internal transporters, and storage space availability. Most of these resources might be dominant and the resulting hub location model can consider a lot of alternative seaport configurations with different levels of the installed resources. In this sense, hub location models with generalized capacity constraints are powerful decision tools by considering not only strategic network design decisions but also aid strategic hub design decisions.

In future works we intend to develop exact solution algorithms to solve the hub location problem with generalized capacity constraints, especially in the case where there is a large number of hub configurations. Our results suggest that designed networks may contain hubs with a high utilization, increasing congestion. To prevent over-utilization of hubs, we had used different percentiles of

capacity parameters as a simple approach. We hope to extend the models to include congestion by relying on queuing theory or introducing congestion costs.

Appendix A. Estimation of the expected ITU dwell time considering stochastic batch service times

Let $f_X(x)$ be the probability density function of the batch service time. Let x some realization of the batch service time. Consider the case of an import ITU, the expected ITU dwell time for the realization of the batch service time x is $x/\ln\left(1/(1-r_h^{\mathrm{i}})\right)$ (equation (23)). The expected ITU dwell time considering all possible realizations of the batch service time is

$$\tau_{hq}^{\mathrm{i}} = \int_{-\infty}^{\infty} \frac{x}{\ln\left(\frac{1}{1-r_h^{\mathrm{i}}}\right)} f_X(x)dx = \frac{1}{\ln\left(\frac{1}{1-r_h^{\mathrm{i}}}\right)} \int_{-\infty}^{\infty} x f_X(x)dx = \frac{T_{hq}}{\ln\left(\frac{1}{1-r_h^{\mathrm{i}}}\right)}. \tag{A.1}$$

The same reasoning is applied to export ITUs. Note that the randomness of the fractions of direct transshipments does not have any impact on the estimation of the ITU dwell time since the expected fractions of direct transshipments r_h^{e} and r_h^{i} are taken as a parameter in equations (23) and (25), and not as a random variable.

Appendix B. Results for scenario 3

Table B.2 shows the results for scenario 3 under the COL data set. The first column of the table gives the instance name. The third column shows the number of hub configurations $|Q|$. The next two columns under the heading *LPR Gap (%)* show the LP relaxation gap for the `BM` and `MC` formulations, respectively. The next two columns under the heading *B&B nodes* show the number of branch-and-bound nodes evaluated at the enumeration tree. The columns under the heading *Total time (sec)* show the total time in seconds for each formulation. The columns under the heading *Obj. val.* show the result of the objective value in the optimal solution. Finally, the column under heading *Opt. hubs - Hub Conf.* shows the optimal hubs along with the optimal hub configuration. The term *Time* is used when the CPLEX time limit is reached.

Table B.2: Results for the limited storage space availability scenario.

		LPR gap (%)		B&B nodes		Time (seconds)			
Instance	$\|Q\|$	BM	MC	BM	MC	BM	MC	Obj. val.	Opt. hubs - Hub configuration
10-1	8	3.32	3.32	101	90	4.7	1.2	419,012.96	1-4,2-4,3-4,4-3,7-4,8-3,10-4
10-2	9	3.22	3.22	197	371	6.2	1.6	419,903.48	1-5,2-5,3-7,4-4,7-4,8-4,10-5
10-3	11	2.86	2.89	232	889	7.6	2.9	419,903.48	1-6,2-6,3-9,4-5,7-5,8-5,10-6
10-4	14	2.58	2.60	269	1,025	15.1	4.2	420,095.44	1-6,2-7,3-10,4-5,7-5,8-5,10-7
10-5	16	2.41	2.40	1,078	3,263	17.7	5.0	420,799.71	1-6,2-7,3-10,4-5,7-5,8-5,10-7
10-6	16	2.28	2.28	1,048	2,359	18.6	4.7	421,809.04	1-7,2-7,3-10,4-6,7-6,8-5,10-7
10-7	16	2.17	2.17	1,085	3,585	19.8	5.6	422,971.09	1-8,2-7,3-11,4-7,7-7,8-5,10-7
10-8	16	2.00	2.00	493	1,839	22.2	3.3	423,904.62	1-8,2-7,3-11,4-7,7-7,8-6,10-7
10-9	12	1.83	1.83	144	488	10.7	2.1	425,559.33	1-6,2-7,3-9,4-5,7-6,8-4,10-6
20-1	8	2.75	2.75	2,587	3,406	227.5	216.3	458,307.55	1-4,2-4,3-4,4-3,13-2,14-4,15-2
20-2	9	2.84	2.84	3,571	6,526	498.3	557.8	459,960.61	1-5,2-5,3-5,4-4,13-2,14-5
20-3	11	2.82	2.83	6,273	9,362	1,891.3	1,542.5	461,138.29	1-6,2-6,3-6,4-5,14-6,15-5
20-4	14	2.69	2.70	9,131	12,468	3,005.9	1,886.4	461,798.86	1-6,2-7,3-9,4-5,14-7,15-5
20-5	16	2.51	2.52	6,570	6,683	2,761.3	846.4	462,215.12	1-6,2-7,3-10,4-5,14-7,15-5
20-6	16	2.40	2.40	5,262	6,561	1,954.5	1,379.6	462,969.87	1-7,2-7,3-10,4-5,14-7,15-5
20-7	16	2.34	2.34	5,156	6,328	3,062.9	825.7	464,030.29	1-6,2-7,3-10,4-5,14-6,15-5,18-7
20-8	16	2.29	2.25	11,027	8,335	3,357.0	1,325.6	465,044.22	1-6,2-7,3-11,4-6,14-6,15-6,18-7
20-9	12	2.08	2.08	4,167	4,455	649.8	218.9	466,242.02	1-6,2-6,3-9,4-5,14-7,15-4
25-1	8	2.64	2.64	3,875	3,475	1,514.0	1,412.6	468,975.62	1-4,3-4,4-4,5-3,17-2,18-4

Table B.2 continued

		LPR gap (%)		B&B nodes		Time (seconds)			
Instance	$\|Q\|$	BM	MC	BM	MC	BM	MC	Obj. val.	Opt. hubs - Hub configuration
25-2	9	2.70	2.70	6,423	2,610	5,958.3	1,618.7	470,432.30	1-5,3-5,4-5,5-4,17-2,18-5
25-3	11	2.69	2.70	9,810	4,412	9,045.1	2,230.0	471,626.77	1-6,3-6,4-6,5-5,18-6,19-5
25-4	14	2.61	2.62	10,678	9,175	11,233.1	5,583.2	472,482.25	1-6,3-7,4-9,5-5,18-7,19-5
25-5	16	2.44	2.45	10,140	9,987	12,401.9	6,729.9	472,914.23	1-6,3-7,4-10,5-5,18-7,19-5
25-6	16	2.33	2.33	6,761	7,755	9,742.7	6,434.7	473,679.81	1-7,3-7,4-10,5-5,18-7,19-5
25-7	16	2.29	2.29	8,985	6,504	10,534.7	4,133.0	474,793.77	1-6,3-7,4-10,5-5,18-6,19-5,22-7
25-8	16	2.24	2.19	8,576	10,549	8,547.5	6,047.7	475,707.26	1-6,3-7,4-10,5-6,18-6,19-6,22-7
25-9	12	2.05	2.04	4,827	6,925	2,703.5	4,309.7	477,014.02	1-4,3-6,4-9,5-5,18-7,19-4
40-1	8	2.09	2.09	1,583	2,162	43,744.5	33,366.0	472,731.88	1-4,4-3,19-4,29-2,31-4,34-2
40-2	9	2.11	2.11	2,118	2,566	Time	45,535.5	474,038.16	1-5,4-4,19-5,29-2,31-5,34-4
40-3	11	2.12	2.12	2,478	4,528	Time	Time	475,276.89	1-6,4-5,19-6,29-3,31-6,34-5
40-4	14	2.18	2.18	10,337	21,050	Time	81,729.4	476,751.71	1-8,4-6,19-7,29-3,31-7,34-5
Average		2.45	2.45	4,677	5,475	12,656.0	9,496.0		

Acknowledgments

This work was supported by Fondo de Ciencia Tecnología e Innovación of Sistema General de Regalías (FCTeI-SGR) of Colombia and Ministerio de Ciencia, Tecnología e Innovación (MINCIENCIAS) of Colombia; and by the Universidad del Valle, Cali, Colombia.

Declarations of Interest

No potential conflict of interest was reported by the authors.

References

Alumur, S. (2009). *Hub location and hub network design*. Ph.D. thesis Bilkent University.

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

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

Alumur, S., Nickel, S., Saldanha-da Gama, F., & Seçerdin, Y. (2016). Multi-period hub network design problems with modular capacities. *Annals of Operations Research*, *246*, 289–312.

Arnold, P., Peeters, D., & Thomas, I. (2004). Modelling a rail/road intermodal transportation system. *Transportation Research Part E: Logistics and Transportation Review*, *40*, 255–270.

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

Balas, E. (1998). Disjunctive programming: Properties of the convex hull of feasible points. *Discrete Applied Mathematics*, *89*, 3–44.

Basallo-Triana, M., Bravo-Bastidas, J., & Vidal-Holguín, C. (2022). A rail-road transshipment yard picture. *Transportation Research Part E: Logistics and Transportation Review*, *159*. doi:10.1016/j.tre.2022.102629.

Basallo-Triana, M., Vidal-Holguín, C., & Bravo-Bastidas, J. (2021). Planning and design of intermodal hub networks: A literature review. *Computers and Operations Research*, *136*.

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

Bontekoning, Y. (2006). *Hub Exchange Operations in Intermodal Hub-and-spoke networks: Comparison of the Performances of Four Types of Rail-rail Exchange Facilities*. Ph.D. thesis Delft University of Technology.

Bouchery, Y., & Fransoo, J. (2015). Cost, carbon emissions and modal shift in intermodal network design decisions. *International Journal of Production Economics*, *164*, 388–399.

Boysen, N., Fliedner, M., Jaehn, F., & Pesch, E. (2013). A survey on container processing in railway yards. *Transportation Science*, *47*, 312–329.

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

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

Contreras, I., Cordeau, J.-F., & Laporte, G. (2012). Exact solution of large-scale hub location problems with multiple capacity levels. *Transportation Science*, *46*, 439–459.

Contreras, I., & O'Kelly, M. (2019). Hub location problems. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.), *Location Science* (pp. 327–363). Cham: Springer International Publishing.

Corry, P., & Kozan, E. (2006). An assignment model for dynamic load planning of intermodal trains. *Computers and Operations Research*, *33*, 1–17.

Crainic, T., & Kim, K. (2007). Chapter 8 Intermodal Transportation. *Handbooks in Operations Research and Management Science*, *14*, 467–537.

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

Ernst, A., & Krishnamoorthy, M. (1996). Efficient algorithms for the uncapacitated single allocationp-hub median problem. *Location Science*, *4*, 139–154.

Federgruen, A., & Groenevelt, H. (1986). M/g/c queueing systems with multiple customer classes: characterization and control of achievable performance under non-preemptive priority rules. *Management Science*, *34*, 1121–1138.

Fotuhi, F., & Huynh, N. (2015). Intermodal network expansion in a competitive environment with uncertain demands. *International Journal of Industrial Engineering Computations*, *6*, 285–304.

García, A., & García, I. (2012). A simulation-based flexible platform for the design and evaluation of rail service infrastructures. *Simulation Modelling Practice and Theory*, *27*, 31–46.

Ghaffari-Nasab, N., Ghazanfari, M., & Teimoury, E. (2015). Robust optimization approach to the design of hub-and-spoke networks. *International Journal of Advanced Manufacturing Technology*, *76*, 1091–1110.

Ishfaq, R., & Sox, C. (2010). Intermodal logistics: The interplay of financial, operational and service issues. *Transportation Research Part E: Logistics and Transportation Review*, *46*, 926–949.

Ishfaq, R., & Sox, C. (2011). Hub location-allocation in intermodal logistic networks. *European Journal of Operational Research*, *210*, 213–230.

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

Janic, M. (2007). Modelling the full costs of an intermodal and road freight transport network. *Transportation Research Part D: Transport and Environment*, *12*, 33–44.

Jeroslow, R., & Lowe, J. (1983). Modelling with integer variables. *Mathematical Programming Study*, (pp. 167–184).

Kahag, M., Niaki, S., Seifbarghy, M., & Zabihi, S. (2019). Bi-objective optimization of multi-server intermodal hub-location-allocation problem in congested systems: modeling and solution. *Journal of Industrial Engineering International*, *15*, 221–248.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kreutzberger, E., & Konings, R. (2016). The challenge of appropriate hub terminal and hub-and-spoke network development for seaports and intermodal rail transport in europe. *Research in Transportation Business and Management*, *19*, 83–96.

Li, S.-X., Sun, S.-F., Wang, Y.-Q., Wu, Y.-F., & Liu, L.-P. (2017). A two-stage stochastic programming model for rail-truck intermodal network design with uncertain customer demand. *Journal of Interdisciplinary Mathematics*, *20*, 611–621.

Limbourg, S., & Jourquin, B. (2009). Optimal rail-road container terminal locations on the european network. *Transportation Research Part E: Logistics and Transportation Review*, *45*, 551–563.

Lin, C.-C., Chiang, Y.-I., & Lin, S.-W. (2014). Efficient model and heuristic for the intermodal terminal location problem. *Computers and Operations Research*, *51*, 41–51.

Lin, C.-C., & Lin, S.-W. (2016). Two-stage approach to the intermodal terminal location problem. *Computers and Operations Research*, *67*, 113–119.

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

Marufuzzaman, M., Eksioglu, S., Li, X., & Wang, J. (2014). Analyzing the impact of intermodal-related risk to the design and management of biofuel supply chain. *Transportation Research Part E: Logistics and Transportation Review*, *69*, 122–145.

Meng, Q., & Wang, X. (2011). Intermodal hub-and-spoke network design: Incorporating multiple stakeholders and multi-type containers. *Transportation Research Part B: Methodological*, *45*, 724–742.

Merakli, M., & Yaman, H. (2016). Robust intermodal hub location under polyhedral demand uncertainty. *Transportation Research Part B: Methodological*, *86*, 66–85.

Mohammadi, M., Jula, P., & Tavakkoli-Moghaddam, R. (2019). Reliable single-allocation hub location problem with disruptions. *Transportation Research Part E: Logistics and Transportation Review*, *123*, 90–120.

Najy, W., & Diabat, A. (2020). Benders decomposition for multiple-allocation hub-and-spoke network design with economies of scale and node congestion. *Transportation Research Part B: Methodological*, *133*, 62–84.

Racunica, I., & Wynter, L. (2005). Optimal location of intermodal freight hubs. *Transportation Research Part B: Methodological*, *39*, 453–477.

Rizzoli, A., Fornara, N., & Gambardella, L. (2002). A simulation tool for combined rail/road transport in intermodal terminals. *Mathematics and Computers in Simulation*, *59*, 57–71.

Rotter, H. (2004). New operating concepts for intermodal transport: The mega hub in hanover/lehrte in germany. *Transportation Planning and Technology*, *27*, 347–365.

Saeedi, H., Behdani, B., Wiegmans, B., & Zuidwijk, R. (2019). Assessing the technical efficiency of intermodal freight transport chains using a modified network DEA approach. *Transportation Research Part E*, *126*, 66–86.

Serper, E., & Alumur, S. (2016). The design of capacitated intermodal hub networks with different vehicle types. *Transportation Research Part B: Methodological*, *86*, 51–65.

Shortle, J., Thompson, J., Gross, D., & Harris, C. (2018). *Fundamentals of Queueing Theory*. Wiley Series in Probability and Statistics. New York: Wiley.

Sörensen, K., & Vanovermeire, C. (2013). Bi-objective optimization of the intermodal terminal location problem as a policy-support tool. *Computers in Industry*, *64*, 128–135.

Steadieseifi, M., Dellaert, N., Nuijten, W., Van Woensel, T., & Raoufi, R. (2014). Multimodal freight transportation planning: A literature review. *European Journal of Operational Research*, *233*, 1–15.

Taherkhani, G., & Alumur, S. (2019). Profit maximizing hub location problems. *Omega*, *86*, 1–15.

Taherkhani, G., Alumur, S., & Hosseini, M. (2020). Benders decomposition for the profit maximizing capacitated hub location problem with multiple demand classes. *Transportation Science*, *54*, 1446–1470.

Teye, C., Bell, M., & Bliemer, M. (2017). Urban intermodal terminals: The entropy maximising facility location problem. *Transportation Research Part B: Methodological*, *100*, 64–81.

Teye, C., Bell, M., & Bliemer, M. (2018). Locating urban and regional container terminals in a competitive environment: An entropy maximising approach. *Transportation Research Part B: Methodological*, *117*, 971–985.

Vielma, J. (2015). Mixed integer linear programming formulation techniques. *SIAM Review*, *57*, 3–57.

Wang, R., Yang, K., Yang, L., & Gao, Z. (2018). Modeling and optimization of a rail-road intermodal transport system under uncertain information. *Engineering Applications of Artificial Intelligence*, *72*, 423–436.

Wang, X., & Meng, Q. (2017). Discrete intermodal freight transportation network design with route choice behavior of intermodal operators. *Transportation Research Part B: Methodological*, *95*, 76–104.

Wiegmans, B., & Behdani, B. (2018). A review and analysis of the investment in, and cost structure of, intermodal rail terminals. *Transport Reviews*, *38*, 33–51.

Yang, K., Yang, L., & Gao, Z. (2016). Planning and optimization of intermodal hub-and-spoke network under mixed uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, *95*, 248–266.

Zhalechian, M., Tavakkoli-Moghaddam, R., & Rahimi, Y. (2017a). A self-adaptive evolutionary algorithm for a fuzzy multi-objective hub location problem: An integration of responsiveness and social responsibility. *Engineering Applications of Artificial Intelligence*, *62*, 1–16.

Zhalechian, M., Tavakkoli-Moghaddam, R., Rahimi, Y., & Jolai, F. (2017b). An interactive possibilistic programming approach for a multi-objective hub location problem: Economic and environmental design. *Applied Soft Computing Journal*, *52*, 699–713.