Submitted to *Transportation Science*
manuscript (Please, provide the manuscript number!)

# Intermodal hub network design with probabilistic service level constraints

**(Authors' names blinded for peer review)**

The design of intermodal hub networks often requires the satisfaction of service time requirements related to the transportation times on the arcs of the network and the processing times at the hub nodes. In this study, we characterize the probability distribution of the total service time in intermodal hub networks and introduce associated probabilistic service level constraints. The resulting formulation incorporates a substantial number of non-linear logical constraints involving indicator variables. Probabilistic service level constraints are formulated in the form of perspective cuts, which are introduced in a cutting plane framework. Additionally, we propose an approximated formulation along with an extended formulation and valid inequalities for the problem. The approximated formulation produces high-quality solutions and is solved more efficiently. For its part, the extended exact formulation can solve a larger number of instances and requires a smaller number of cuts to be added during the cutting plane process. Extensive computational experiments are performed on the Australian Post (AP) and Colombian (COL) datasets to evaluate the efficiency and limitations of the proposed formulations and solution algorithms. Results obtained on 480 problem instances under different service level scenarios confirm the effectiveness of the proposed formulations. We also provide managerial insights based on detailed sensitivity analyses to assess the effect of varying service level requirements on the optimal network configurations.

*Key words*: Hub network design, intermodal transportation, probabilistic service level constraints, cutting plane algorithm, hypo-exponential distribution.

## 1. Introduction

Intermodal transportation plays a crucial role in satisfying the demand for shipments in many-to-many distribution networks. Under this concept, two or more transportation modes are used alongside specialized facilities or hubs to consolidate shipments and transship goods using different modes. Inter-hub transportation involves high-capacity vehicles, such as trains or ships, leveraging economies of scale in transportation operations, making intermodal transportation attractive in practice. Some reviews on intermodal transportation include Crainic and Kim (2007), Steadieseifi et al. (2014), and Basallo-Triana et al. (2021).

Despite the economic benefits of intermodal transportation, the quality of service perceived by customers is usually impacted by longer transit times. Intermodal transportation generally results in longer transportation times compared to direct transportation. Congestion at hubs is another significant issue. For example, only 53% of freight trains reach their destination with fewer than 30 minutes delay (Boysen et al. 2013). Service times are notably increased predominantly due to long waiting times at hubs. A customer service level of this magnitude drastically reduces competitiveness in intermodal transportation.

This work aims to design intermodal hub networks considering the quality of service, which is measured as the probability of satisfying a service time requirement for transportation from origin to destination. The total service time is the primary determinant of the network's service quality. This comprises transportation time as well as waiting and processing times at hubs. The service time encompasses various sources of uncertainty, making it imperative to adopt a holistic stochastic approach to simultaneously integrate all sources of uncertainty into a hub network design model. The design of intermodal hub networks seeks to determine the optimal number and locations of hubs, the routing of flows from origin to destination nodes through the hubs, and direct transportation. Recent surveys on hub location include Contreras and O'Kelly (2019), and Alumur et al. (2021).

Uncertainty in the hub network design literature is usually addressed as robust or stochastic models. Contributions to the robust optimization approach include Zetina et al. (2017), Martins de Sá et al. (2018a), and Martins de Sá et al. (2018b), by considering uncertainty intervals for different parameters. For their part, Merakli and Yaman (2016), Ghaffarinasab (2018) and Shahabi and Unnikrishnan (2014) introduced more general uncertainty sets.

Some stochastic programming contributions utilizing the expected value criterion are Contreras et al. (2011), Alumur et al. (2012), Taherkhani et al. (2020), and Taherkhani et al. (2021). Most formulations employ sampling methods to estimate the expectations. In contrast, Ghaffarinasab (2022) provided a mathematical expression for the expected value of Bernoulli-distributed demands. The expected value approach has also been extensively applied in congestion-related formulations through queueing models. Studies in this direction include Mohammadi et al. (2017), Azizi et al. (2018), and Ishfaq and Sox (2012). More recently, some authors have suggested a measure related to the variability of random variables instead of the expected value criterion (Ghaffarinasab et al. 2023).

Chance-constrained formulations for normally distributed parameters include Gao and Qin (2016), Mohammadi et al. (2013), Hu et al. (2021). For their part Marianov and Serra (2003) and Mohammadi et al. (2011) propose probabilistic (chance) constraints for the waiting time at hubs, which is estimated using multiserver queuing models. Jayaswal and Vidyarthi (2023) introduced a hub location model with probabilistic service level constraints on the sojourn time hubs considering a heterogeneous, preemptive priority M/M/1 queuing model with two customer classes.

We introduce the *intermodal hub network design problem with probabilistic service level constraints* (IHNDP-SLC) on the intermodal transportation route. Probabilistic constraints ensure that commodities using the intermodal routes are delivered within a prescribed service time limit, with a probability greater than or equal to a threshold value. The service time includes the transportation time, and the waiting and processing times at hubs, which are random variables.

The contributions of this paper are threefold. First, we describe the theoretical properties of the total service time distribution function as the convolution of the total transport time density function and the total sojourn time distribution at hubs. Hubs are modeled as M/M/1 queuing systems, and the transportation time is assumed to follow an arbitrary probability density function.

Second, we exploit the theoretical properties of the total sojourn time distribution to simplify the structure of probabilistic service level constraints, which are formulated as nonlinear constraints involving indicator variables. We conjecture the convexity of the probabilistic service level constraints and validate this conjecture through several numerical experiments. We use perspective cuts to reformulate the logical conditions imposed by indicator variables, which are added as cutting planes during the execution of the branch-and-bound algorithm. The resulting optimization problem is highly challenging to solve because of the large number of indicator variables and perspective cuts.

Third, we propose a method for obtaining a tractable approximation of probabilistic service level constraints using a family of homothetic functions. The accuracy of the approximations is notably high and the resulting approximation allows us to considerably reduce the number of cutting planes needed to solve the problem. The approximated formulation relies on less accurate (coarse) perspective cuts and is solved between two and seven times faster than the exact formulation. Finally, we propose an extended formulation for the exact problem by introducing coarse perspective cuts as valid inequalities. The extended formulation is solved between one and two times faster than our initial exact formulation. The effectiveness of the approximated and extended formulations is increased as the network size increases.

We report the results of extensive computational experiments based on the Australian Post (AP) dataset and Colombian (COL) dataset to evaluate the efficiency and limitations of the proposed formulations and solution algorithms. We also provide managerial insights based on detailed sensitivity analyses to assess the effect of varying service level requirements on the optimal intermodal hub network configurations.

The remainder of the paper is organized as follows: Section 2 presents the basic intermodal hub network design model and introduces probabilistic service level constraints. Sections 3 and 4 describe theoretical properties of the total sojourn time distribution and the total service time distribution, respectively. Sections 5 and 6 present mixed integer linear programming formulations and cutting plane algorithms for the hub network design problem with probabilistic service level constraints. Section 7 discusses the results of numerical experiments. Finally, Section 8 concludes the paper.

## 2. Model Formulation

In this section, we present a basic formulation of the intermodal hub network design problem and introduce the probabilistic service level constraints.

### 2.1. Basic intermodal hub network design formulation

Let $G=(N,A)$ be a complete digraph where $N$ is the set of nodes and $A$ is the set of directed arcs. A hub arc is an ordered set $a=\{a_1,a_2\}\in A$, where $a_1$ and $a_2$ $(a_1\neq a_2)$ are potential hub locations. Let $K$ be the set of commodities or origin-destination (OD) pairs with strictly positive demand: $K=\left\{(i,j)\in N\times N \mid w_{(i,j)}>0, i\neq j\right\}$. We consider both direct (truck only) and intermodal transportation (truck-train-truck) options. In direct transportation, freight is transported by trucks without using the intermediate hubs. Intermodal transportation involves moving freight using a combination of trucks and trains, requiring transshipment at hubs. Hence, the demand for commodity $k\in K$ can be satisfied partially or completely using the intermodal transportation option. The flow that is not routed through the intermodal option is transported using the direct transportation option. We consider that intermodal transportation uses exactly two hubs, and there are no single hub paths in the network.

Let $c_{ij}$ be the distance between nodes $i\in N$ and $j\in N$. The total intermodal transportation cost of commodity $k\in K$ through hub arc $a\in A$ is given by $C_{ka}=C_{(i,j)(a_1,a_2)}=w_k(\chi c_{ia_1}+\gamma c_{a_1a_2}+\delta c_{a_2j})$, where $\chi,\gamma$ and $\delta$ are the unit collection, transfer and distribution costs, respectively. To account for the economies of scale on the inter-hub link there is a discount factor such that $\gamma<\chi$ and $\gamma<\delta$. The total direct transportation cost of commodity $k\in K$ is denoted by $C_k=C_{(i,j)}=\rho w_k c_k$, where $\rho$ is the unit transportation cost for direct shipment. We assume that $\rho>\chi$ and $\rho>\delta$. For every commodity $k\in K$, the most economical transportation route for satisfying the demand is always considered (see Taherkhani et al. 2020). Hence, the set of relevant hub arcs for commodity $k$ is defined as follows:

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

Let $z_m\in\{0,1\}$ be a binary decision variable that equals 1 if a hub facility is located at node $m\in N$, and 0 otherwise. The fixed cost of opening a hub facility at node $m\in N$ is denoted by $f_m$. Let $\Lambda_m$ denote the capacity of hub $m$. We will explain later the specific role of this parameter in the context of probabilistic service level constraints. Let $x_{ka}$ be the fraction of demand of commodity $k \in K$ that is satisfied using the intermodal transportation option through hub arc $a \in A_k$. In that case, the fraction of demand of commodity $k \in K$ that is satisfied using the direct shipment is denoted by $1 - \sum_{a \in A_k} x_{ka}$. The formulation of the basic intermodal (capacitated) hub network design model is as follows:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{m \in N} f_m z_m + \sum_{k \in K} \sum_{a \in A_k} C_{ka} x_{ka} + \sum_{k \in K} C_k \left( 1 - \sum_{a \in A_k} x_{ka} \right) && \text{(2a)} \\
\text{subject to} \quad & \sum_{a \in A_k} x_{ka} \leq 1, && \forall k \in K, && \text{(2b)} \\
& \sum_{\substack{a \in A_k, \\ m \in a}} x_{ka} \leq z_m, && \forall m \in N, k \in K, && \text{(2c)} \\
& \sum_{k \in K} \sum_{\substack{a \in A_k, \\ m \in a}} w_k x_{ka} \leq \Lambda_m z_m, && \forall m \in N, && \text{(2d)} \\
& x_{ka} \in [0,1], && \forall k \in K, a \in A_k && \text{(2e)} \\
& z_m \in \{0,1\}, && \forall m \in N. && \text{(2f)}
\end{aligned}
$$

The objective function (2a) minimizes the sum of hub location and transportation costs. The first term in the objective function is the total fixed cost of installing hub facilities. The second term captures the total intermodal transportation cost, and the last term corresponds to the total direct (truck-only) transportation cost. Note that the objective function can be equivalently written as:

$$
\sum_{m \in N} f_m z_m - \sum_{k \in K} \sum_{a \in A_k} (C_k - C_{ka}) x_{ka} + \sum_{k \in K} C_k, \tag{3}
$$

where $C_k - C_{ka}$ denotes the saving achieved by using intermodal transportation instead of the direct transportation option for satisfying the demand of commodity $k \in K$ using hub arc $a \in A_k$. Note that, by the definition of set $A_k$ in equation (1), we have $C_k - C_{ka} > 0$, for all $k \in K$ and $a \in A_k$.

Constraints (2b) ensure that the demand of commodity $k \in K$ is fully or partially satisfied using the intermodal transportation option. For a commodity, if constraints (2b) are binding at optimality, then the commodity is routed through intermodal transportation only. In that case, the direct transportation option is discarded for such a commodity. Constraints (2c) prohibit flows from being routed through a non-open hub. Constraints (2d) limit the total amount of flow that can be processed at a hub. Finally, Constraints (2e)-(2f) define the domain of the decision variables.

Formulation (2) does not account for the possibility that high utilization of the available hub capacity could lead to congestion and, consequently, an increased waiting time at the hubs. Additionally, transportation consume a significant amount of time. These factors are pivotal in determining the quality of service offered by the intermodal hub network. It is worthwhile to incorporate a quality of service into the design of intermodal hub networks.

### 2.2. Probabilistic service level constraints

The congestion at hubs may cause the delay of shipments resulting in missed promised delivery times, penalties, or an expensive expedited delivery (to avoid any further delay and/or penalties) (Jayaswal and Vidyarthi 2023). Hence, the transportation service provider should set its maximum service time and a target service level as measures of service quality. The *total service time* of the intermodal transportation comprises the transportation time on the collection arc, inter-hub arc, and distribution arc as well as the waiting and processing times at hub facilities. Figure 1 illustrates the components of the total service time in a hub network setting where the flow for commodity $k=(i,j)$ passes through exactly two hubs. The *sojourn time* at a hub represents the total time the flow units spend in waiting and service at the hub facility. The sojourn times at hubs $a_1$ and $a_2$ are denoted by $V_{a_1}$ and $V_{a_2}$, respectively. The total sojourn time for a given commodity using the hub arc $a=(a_1,a_2)\in A$ is denoted by $V_a = V_{a_1}+V_{a_2}$. The collection time $U^1_{ka}$ is defined as the transportation time from origin node $i$ to hub node $a_1$. The transfer time $U^2_{ka}$ denotes the time taken for inter-hub transportation between hub nodes $a_1$ and $a_2$. The distribution time $U^3_{ka}$ represents the transportation time between hub node $a_2$ and destination node $j$. The total transportation time is denoted by $U_{ka}=U^1_{ka}+U^2_{ka}+U^3_{ka}$.

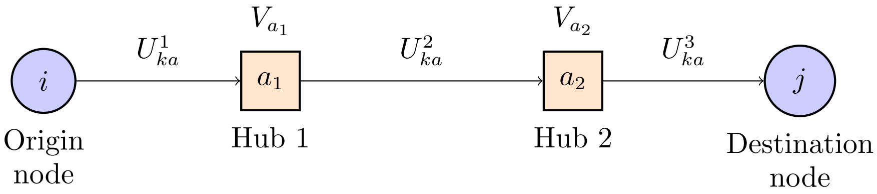


**Figure 1 The elements of the total service time of commodity** $k=(i,j)\in K$ **through hub arc** $a=(a_1,a_2)\in A_k$**.**

Let $\lambda_{a_1}$ and $\lambda_{a_2}$ denote the total flow that passes through hubs $a_1$ and $a_2$, respectively. The total flow at hub node $a_1$ and $a_2$ are defined in terms of decision variables $x_{ka}$ as in (4). It is important to note that this flow includes both the collection and transfer flows through the respective hub.

$$\lambda_{a_1}=\sum_{k\in K}\sum_{\substack{a\in A_k,\\ a_1\in a}} w_k x_{ka}, \quad \forall a_1\in N, \qquad \lambda_{a_2}=\sum_{k\in K}\sum_{\substack{a\in A_k,\\ a_2\in a}} w_k x_{ka}, \qquad \forall a_2\in N. \tag{4}$$

Let $V_a$ and $U_{ka}$ be random variables, the total service time $V_a+U_{ka}$ is also a random variable with a distribution function denoted as $S_{T_{ka}}$. If the maximum acceptable service time to fulfill the demand of commodity $k$ is $\tau_k>0$, then the service level is defined as the probability that the total service time $V_a+U_{ka}$ does not exceed the maximum allowable service time $\tau_k$. If $\alpha\in[0,1]$ denotes the (minimum) target service level, then the *probabilistic service level constraints* can be expressed as follows:

$$S_{T_{ka}}(\tau_k|\lambda_{a_1},\lambda_{a_2})=P\left(V_a+U_{ka}\le\tau_k|\lambda_{a_1},\lambda_{a_2}\right)\ge\alpha, \quad \text{for } x_{ka}>0, k\in K, a\in A_k. \tag{5}$$

Note that these constraints are relevant only in the case when commodity $k$ uses hub arc $a$ to satisfy its demand, i.e., $x_{ka} > 0$. We also emphasize that the probability on the left-hand side of (5) depends on the total flow $\lambda_{a_1}$ and $\lambda_{a_2}$ that passes through hubs $a_1$ and $a_2$, respectively. This is because the sojourn time in a hub depends on the level of congestion at the hub. We are interested in analyzing the probability distribution of the total service time defined on the left-hand side of (5). The fact that constraint (5) should be activated only if $x_{ka} > 0$, translates into a logical condition that can be modeled as nonlinear disjunctive inequalities, rendering it challenging to solve.

We introduce the notation in Table 1 for the probabilistic analysis.

Table 1: Notation used for the probabilistic analysis

| | |
|---|---|
| *Random variables* | |
| $U_{ka}$ | Total transportation time, including the collection, transfer, and distribution stages, $k \in K, a \in A_k$. |
| $V_m$ | Sojourn time at hub $m \in N$. |
| $V_a$ | Total sojourn time $V_a = V_{a_1} + V_{a_2}, (a_1, a_2) \in A$. |
| $T_{ka}$ | Total service time for commodity $k \in K$ and hub arc $a \in A_k$, $T_{ka} = V_a + U_{ka}$. |
| *Densities and distribution functions* | |
| $w_{V_m}(v_m \mid \lambda_m)$ | PDF for the sojourn time (including waiting and service time) at hub $m \in N$. |
| $W_{V_m}(v_m \mid \lambda_m)$ | CDF for the sojourn time (including waiting and service time) at hub $m \in N$. |
| $W_{V_a}(v_a \mid \lambda_{a_1}, \lambda_{a_2})$ | CDF for the total sojourn time (including waiting and service times) on hub arc $a = (a_1, a_2) \in A$. Note that $V_a = V_{a_1} + V_{a_2}$ and $v_a = v_{a_1} + v_{a_2}$. |
| $g_{U_{ka}}(u_{ka})$ | PDF of the total transport time for commodity $k \in K$ at inter-hub link $a \in A_k$. |
| $G_{U_{ka}}(u_{ka})$ | CDF of the total transport time for commodity $k \in K$ at inter-hub link $a \in A_k$. |
| $S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2})$ | CDF of the total service time for commodity $k \in K$ and inter-hub link $a = (a_1, a_2) \in A_k$ evaluated at the maximum allowed service time $\tau_k$. |
| *Parameters of the probabilistic service level constraints* | |
| $\tau_k$ | Maximum accpetable service time for commodity $k \in K$. |
| $\alpha$ | Target (minimum) service level. |
| $\mu_m$ | Processing rate at hub $m \in N$. |
| *Decision variables involved in the probabilistic service level constraints* | |
| $\lambda_m$ | Total flow that passes through hub $m \in N$. |
| *Level sets* | |
| $C_\alpha^{W_\tau}(a)$ | The $\alpha$-super level set of the total sojourn time distribution evaluated at time $\tau$ concerning inter-hub link $a$. |
| $C_\alpha^{S_{\tau_k}}(a)$ | The $\alpha$-super level set of the total service time distribution concerning commodity $k$ and inter-hub link $a$. |
| $E_\alpha^{W_\tau}(a)$ | The $\alpha$ level set of the total sojourn time distribution concerning inter-hub link $a$. |
| $E_\alpha^{S_{\tau_k}}(a)$ | The $\alpha$ level set of the total service time distribution of commodity $k$ and inter-hub link $a$. |

## 3. Analysis of the total sojourn time distribution

If we assume that the flow rates between different origin node-destination node pairs $(i,j)$ are independent random variables that follow a Poisson process with a mean $w_k$, then the aggregate flow rate through hub $m$ also follows a Poisson process with a mean given by $\lambda_m = \sum_{k\in K}\sum_{\substack{a\in A_k,\\ m\in a}} w_k x_{ka}$. Further, we assume that the service times at the hub follow an exponential distribution with a mean of $1/\mu_m$. Then, the hub can be modeled as an M/M/1 queuing system with flexible service capacity $\mu_m$. The queuing system's stability condition requires $0 \leq \lambda_m < \mu_m$. The sojourn time is a random variable, that has an exponential distribution, with the probability density function (PDF) defined by $w_{V_m}(v_m|\lambda_m) = (\mu_m - \lambda_m)e^{-(\mu_m-\lambda_m)v_m}$. The PDF of the total sojourn time on hub arc $a = (a_1, a_2) \in A$ is the convolution of the densities for the sojourn time at hubs $a_1$ and $a_2$: $w_{V_{a_1}} * w_{V_{a_2}}(v_a|\lambda_{a_1}, \lambda_{a_2})$, where the expression "$*$" refers to the convolution operation. For exponential density functions, it is known that such a convolution corresponds to a hypo-exponential density with parameters $(\mu_{a_1} - \lambda_{a_1})$ and $(\mu_{a_2} - \lambda_{a_2})$ (Ross 2010, p. 308). The corresponding cumulative distribution function (CDF) of the total sojourn time (hypo-exponential distribution) on hub arc $a$ is given by:

$$\begin{aligned} W_{V_a}(v_a|\lambda_{a_1}, \lambda_{a_2}) &= W_{V_{a_1}} * w_{V_{a_2}}(v_a|\lambda_{a_1}, \lambda_{a_2}), \\ &= 1 - \frac{(\mu_{a_1} - \lambda_{a_1})e^{-(\mu_{a_2}-\lambda_{a_2})v_a}}{(\mu_{a_1} - \lambda_{a_1}) - (\mu_{a_2} - \lambda_{a_2})} - \frac{(\mu_{a_2} - \lambda_{a_2})e^{-(\mu_{a_1}-\lambda_{a_1})v_a}}{(\mu_{a_2} - \lambda_{a_2}) - (\mu_{a_1} - \lambda_{a_1})}, \end{aligned} \tag{6}$$

where $v_a \geq 0$. Recall that the average processing rates at each hub, $\mu_{a_1}$ and $\mu_{a_2}$, are parameters of the model. However, the average demand rates $\lambda_{a_1}$ and $\lambda_{a_2}$ are decision variables as they are functions of decision variables $(x_{ka})$ as in Equation (4).

It should be noted that Equation (6) has singularities (points where the evaluation of function $W_{V_a}$ is indeterminate) at the line $\mu_{a_1} - \lambda_{a_1} = \mu_{a_2} - \lambda_{a_2}$. To overcome these indeterminate computations, we exploit the fact that when $\mu_{a_1} - \lambda_{a_1} = \mu_{a_2} - \lambda_{a_2}$, the corresponding convolution produces an Erlang density function (Ibe 2009, p. 22). Hence, we can rewrite the expression for the total sojourn time distribution as follows:

$$W_{V_a}(v_a|\lambda_{a_1}, \lambda_{a_2}) = \begin{cases} 1 - \frac{(\mu_{a_1}-\lambda_{a_1})e^{-(\mu_{a_2}-\lambda_{a_2})v_a}}{(\mu_{a_1}-\lambda_{a_1})-(\mu_{a_2}-\lambda_{a_2})} - \frac{(\mu_{a_2}-\lambda_{a_2})e^{-(\mu_{a_1}-\lambda_{a_1})v_a}}{(\mu_{a_2}-\lambda_{a_2})-(\mu_{a_1}-\lambda_{a_1})}, & \text{for } \mu_{a_1} - \lambda_{a_1} \neq \mu_{a_2} - \lambda_{a_2}, \\ 1 - e^{-(\mu_{a_1}-\lambda_{a_1})v_a} - (\mu_{a_1} - \lambda_{a_1})v_a e^{-(\mu_{a_1}-\lambda_{a_1})v_a}, & \text{for } \mu_{a_1} - \lambda_{a_1} = \mu_{a_2} - \lambda_{a_2}. \end{cases} \tag{7}$$

The relevant domain of the total sojourn time distribution as a function of flows (decision variables) is as follows:

$$D = \{(\lambda_{a_1}, \lambda_{a_2}) \mid 0 \leq \lambda_{a_1} < \mu_{a_1}, 0 \leq \lambda_{a_2} < \mu_{a_2}\}. \tag{8}$$

PROPOSITION 1. *The total sojourn time distribution $W_{V_a}(v_a \mid \lambda_{a_1}, \lambda_{a_2})$ defined for $(\lambda_{a_1}, \lambda_{a_2}) \in D$ has the following properties:*

1. *$W_{V_a}(\lambda_{a_1}, \lambda_{a_2})$ is symmetric around the singularity line $\mu_{a_1} - \lambda_{a_1} = \mu_{a_2} - \lambda_{a_2}$.*
2. *$W_{V_a}(\lambda_{a_1}, \lambda_{a_2})$ is continuous and differentiable everywhere on $D$.*
3. *$W_{V_a}(\lambda_{a_1}, \lambda_{a_2})$ is strictly decreasing.*

*Proof* See Appendix A. □

CONJECTURE 1. *Additional properties of the total sojourn time distribution.*

1. *$W_{V_a}(\lambda_{a_1}, \lambda_{a_2})$ is concave at its $100(1-2/e) \approx 26.4$-th percentile or higher.*
2. *$W_{V_a}(\lambda_{a_1}, \lambda_{a_2})$ is quasiconcave.*

Appendix B discusses the concavity of the total sojourn time distribution. We validate numerically the validity of Conjecture 1.

### Levels sets

A useful tool for analyzing probabilistic service level constraints is the concept of level sets.

DEFINITION 1 (LEVEL SETS). The $\alpha$-super level set $C_\alpha^{W_\tau}(a)$ and $\alpha$-level set $E_\alpha^{W_\tau}(a)$ of the total sojourn time distribution evaluated at $v_a = \tau$ are

$$C_\alpha^{W_\tau}(a) := \{(\lambda_{a_1}, \lambda_{a_2}) \mid W_{V_a}(\tau|\lambda_{a_1}, \lambda_{a_2}) \geq \alpha, (\lambda_{a_1}, \lambda_{a_2}) \in D\}, \quad \text{and} \tag{9}$$

$$E_\alpha^{W_\tau}(a) := \{(\lambda_{a_1}, \lambda_{a_2}) \mid W_{V_a}(\tau|\lambda_{a_1}, \lambda_{a_2}) = \alpha, (\lambda_{a_1}, \lambda_{a_2}) \in D\}, \tag{10}$$

respectively.

PROPOSITION 2. *Consider that $E_\alpha^{W_\tau}(a) \neq \emptyset$. If $(\lambda_{a_1}, \lambda_{a_2}) \in E_\alpha^{W_\tau}(a)$, then there exists a bijective relation between $\lambda_{a_1}$ and $\lambda_{a_2}$.*

*Proof* We want to show that if $(\lambda_{a_1}, \lambda_{a_2}) \in E_\alpha^{W_\tau}(a)$, then $\lambda_{a_1}$ is associated to one and only one $\lambda_{a_2}$ in $E_\alpha^{W_\tau}(a)$, and vice-versa. We proced by contradiction. Suppose that $\lambda_{a_1}^1 < \lambda_{a_1}^2$ and that $(\lambda_{a_1}^1, \lambda_{a_2}) \in E_\alpha^{W_\tau}(a)$ and $(\lambda_{a_1}^2, \lambda_{a_2}) \in E_\alpha^{W_\tau}(a)$. This cannot be possible since, by property 3 of Proposition 1, we have $W_{V_a}(\tau|\lambda_{a_1}^1, \lambda_{a_2}) > W_{V_a}(\tau|\lambda_{a_1}^2, \lambda_{a_2})$. □

According to Proposition 2, there exists a function $f^a$ for which $\lambda_{a_1} = f^a(\lambda_{a_2})$. Appendix C derives a closed-form expression for $f^a$, which is shown in Figure 2. In the figure, a blue color represents the case when the branch $r = -1$ of the Lambert W function is chosen, and an orange color when the branch $r = 0$ is chosen. See Appendix C for details.

REMARK 1. Function $f^a$ can be defined as a function of either $\lambda_{a_1}$ or $\lambda_{a_2}$. One convention is to consider $\lambda_{a_1}$ as the dependent variable and $\lambda_{a_2}$ as the independent variable, i.e., $\lambda_{a_1} = f^a(\lambda_{a_2})$, if $a_1 < a_2$. For the definition of $f^a$ the arc direction is irrelevant.

Proposition 3. *An equivalent representation of the $\alpha$-super level set $C^{W_\tau}_\alpha(a)$ is given by:*

$$C^{W_\tau}_\alpha(a)=\{(\lambda_{a_1},\lambda_{a_2})\mid \lambda_{a_1}\le f^a(\lambda_{a_2}),(\lambda_{a_1},\lambda_{a_2})\in D\},\quad \textit{for } a=(a_1,a_2). \tag{11}$$

*Proof* Suppose that $\lambda^1_{a_1}<\lambda^2_{a_1}$ and that $\lambda^2_{a_1}=f(\lambda_{a_2})$, which is equivalent to $(\lambda^2_{a_1},\lambda_{a_2})\in E^{W_\tau}_\alpha(a)$. Given that $(\lambda^2_{a_1},\lambda_{a_2})\in E^{W_\tau}_\alpha a)$ and $E^{W_\tau}_\alpha(a)\subset C^{W_\tau}_\alpha(a)$, we have that $(\lambda^2_{a_1},\lambda_{a_2})\in C^{W_\tau}_\alpha(a)$. Also, by Property 3 in Proposition 1, we have that $W_{V_a}(\tau|\lambda^1_{a_1},\lambda_{a_2})>W_{V_a}(\tau|\lambda^2_{a_1},\lambda_{a_2})=\alpha$, hence, by Definition 1, $(\lambda^1_{a_1},\lambda_{a_2})\in C^{W_\tau}_\alpha(a)$. □

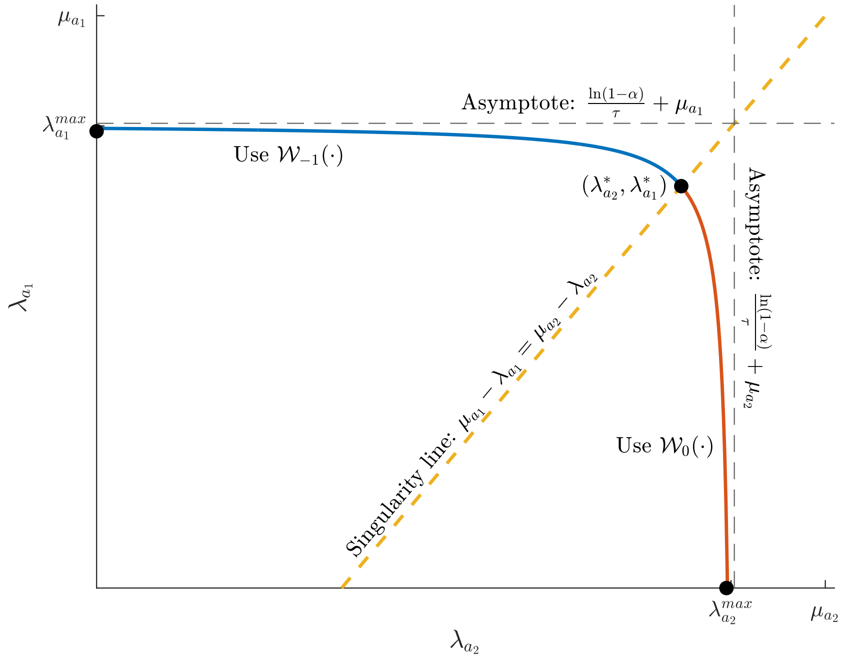


**Figure 2 Plot of function $\lambda_{a_1}=f^a(\lambda_{a_2})$ for inter-hub arc $a=(a_1,a_2)$.**

Proposition 4. *Consider parameters $\tau$ and $\tau'$ such that $0<\tau$ and $0<\tau'\neq\tau$. The $\alpha$-level sets of functions $W_{V_a}(\tau\mid\lambda_{a_1},\lambda_{a_2})$ and $W_{V_a}(\tau'\mid\lambda_{a_1},\lambda_{a_2})$ are homothetic with respect to the point $(\mu_{a_1},\mu_{a_2})$.*

*Proof* See Appendix D □

## 4. Analysis of the total service time distribution

The total service time distribution is the convolution of the total sojourn time distribution and the total transportation time density function. We assume that the probability density function for the total transportation time $g_{U_{ka}}(u_{ka})$ is independent of the flows at each stage of the transportation, and it is symmetric for the transportation path, that is, if $k=(i,j),\bar{k}=(j,i),a=(a_1,a_2),\bar{a}=$

$(a_2, a_1)$, then $g_{U_{ka}}(u_{ka}) = g_{U_{\bar{k}\bar{a}}}(u_{\bar{k}\bar{a}})$. Using the convolution integral, and assuming that the total transportation time and the total sojourn time are independent random variables, the total service time distribution is given by:

$$S_{T_{ka}}(t_{ka} \mid \lambda_{a_1}, \lambda_{a_2}) = W_{Va} * g_{U_{ka}}(t_{ka} \mid \lambda_{a_1}, \lambda_{a_2}) = \int_0^{t_{ka}} W_{Va}(t_{ka} - x \mid \lambda_{a_1}, \lambda_{a_2}) g_{U_{ka}}(x)\, dx, \qquad (12)$$

where $(\lambda_{a_1}, \lambda_{a_2}) \in D$. Given that $g_{U_{ka}}(x) \geq 0$, the total service time distribution $S_{T_{ka}}(t_{ka} \mid \lambda_{a_1}, \lambda_{a_2})$ satisfies properties outlined in Proposition 1. Moreover, if Conjecture 1 is valid, then the total service time distribution is concave in the region where the total sojourn time distribution is concave. This is because concavity is preserved under integration. According to this, probabilistic service level constraints for the total service time in Equation (5) are convex constraints for high values of service levels, i.e., $\alpha > 0.26$.

Analogously to the case of the total sojourn time distribution, it is possible to write the $\alpha$-level and $\alpha$-super level sets, respectively, in terms of a function $f^{ka}$, defined for each commodity $k \in K$ and inter-hub arc $a \in A$:

$$E_\alpha^{S_{\tau_k}}(a) = \left\{ (\lambda_{a_1}, \lambda_{a_2}) \mid \lambda_{a_1} = f^{ka}(\lambda_{a_2}), (\lambda_{a_1}, \lambda_{a_2}) \in D \right\}, \quad \text{for } k \in K, a \in A, \qquad (13)$$

$$C_\alpha^{S_{\tau_k}}(a) = \left\{ (\lambda_{a_1}, \lambda_{a_2}) \mid \lambda_{a_1} \leq f^{ka}(\lambda_{a_2}), (\lambda_{a_1}, \lambda_{a_2}) \in D \right\}, \quad \text{for } k \in K, a \in A. \qquad (14)$$

However, it might be difficult, or impossible, to obtain a closed form expression for $f^{ka}$, as it was done for $f^a$ in Section 3. This is due to the complexity that the total transportation time density $g_{U_{ka}}(u_{ka})$ imposes on the analytical treatment of the convolution integral. In this case, function $f^{ka}$ is computed using numerical methods, for further details see Appendix E.

The set $C_\alpha^{S_{\tau_k}}(a)$ constitutes an alternative way to conceive probabilistic service level constraints (5) as the $\alpha$-super level set of the total service time distribution.

## 5. Probabilistic service level constraints for the total service time

In light of probabilistic service level constraints, certain inter-hub arcs may become infeasible. Such infeasibilities are detected during a pre-processing phase, as described below.

### 5.1. Preprocessing

We define the set of feasible hub arcs for commodity $k$ as follows:

$$\mathscr{A}_k = \left\{ (a_1, a_2) \mid (a_1, a_2) \in A_k, S_{T_{k(a_1,a_2)}}(\tau_k \mid w_k, w_k) > \alpha \right\}, \qquad \forall k \in K. \qquad (15)$$

The set of commodities that are allowed to use hub $m$ is:

$$\mathscr{K}_m = \{ k \mid k \in K, \exists a \in \mathscr{A}_k \text{ such that } m \in a \}, \qquad \forall m \in \mathscr{N}. \qquad (16)$$

The set of feasible hubs is:

$$\mathcal{N} = \left\{ m \mid m \in N, \max_{k\in K,(m,a_2)\in\mathcal{B}_k} \left\{ S_{T_{k(m,a_2)}}(\tau_k \mid w_k, w_k) \right\} > \alpha \right\}. \tag{17}$$

Probabilistic constraints (5) are reformulated according to the $\alpha$-super level set $C_\alpha^{S_{\tau_k}}(a)$ as follows:

$$\lambda_{a_1} \le f^{ka}(\lambda_{a_2}), \qquad \text{if } x_{ka} > 0, k \in K, a = (a_1, a_2) \in \mathcal{A}_k, \tag{18}$$

where the non-negativity of flow variables was already imposed by the non-negativity constraints in Formulation (2). Let

$$\lambda_{ka_1}^{\max}(a_2) = f^{ka}(0), \qquad \forall k \in K, a = (a_1, a_2) \in \mathcal{A}_k, \tag{19}$$

$$\Lambda_m = \max_{k\in K, n\in\mathcal{N}} \left\{ \lambda_{km}^{\max}(n) \right\}, \qquad \forall m \in \mathcal{N}. \tag{20}$$

Note that parameter $\lambda_{ka_1}^{\max}(a_2)$ is an upper bound for $\lambda_{a_1}$ when hub arc $(a_1, a_2)$ is open and commodity $k$ uses such a hub arc. For its part, $\Lambda_m$ is a valid upper bound for $\lambda_m$ regardless of which hub arcs are open and which commodity is considered. From Equation (20), we note that parameter $\Lambda_m$ varies depending on the service level requirement $\alpha$.

### 5.2. Linear reformulation

We introduce auxiliary binary variables to reformulate the logical condition $x_{ka} > 0$ in constraints (18). To keep the problem tractable, it is desirable to introduce a small number of binary variables, and we resort to the symmetry of the probability density function for the total transportation time to this end. Consider that commodity $k = (i, j)$ uses hub arc $a = (a_1, a_2)$ and commodity $\bar{k} = (j, i)$ uses hub arc $\bar{a} = (a_2, a_1)$. Transport paths $i \to a_1 \to a_2 \to j$ and $j \to a_2 \to a_1 \to i$ have the same transport time density function, and consequently the same probabilistic service level constraints. Indeed, inequalities $\lambda_{a_1} \le f^{ka}(\lambda_{a_2})$ and $\lambda_{a_2} \le \mathbf{inv}(f^{ka})(\lambda_{a_1}) = f^{\bar{k}\bar{a}}(\lambda_{a_1})$ produce the same feasible set. This implies that only one auxiliary binary variable is necessary to represent the logical condition in each transport path.

We say that two transport paths are *indistinguishable* in terms of probabilistic service level constraints if one transport path is the reverse of the other and both transport paths have the same transport time density function. Otherwise, we say that the paths are *distinguishable*.

Let $\mathcal{B}$ be the set of undirected hub arcs, and let $\mathcal{L}_o$ be the set of strictly distinguishable paths associated with hub arc $o \in \mathcal{B}$. Let $l \in \mathcal{L}_o$, then $l$ is associated with at most two transport paths, i.e., the path for commodity $k$ using hub arc $a$ and the path for commodity $\bar{k}$ using hub arc $\bar{a}$. Since paths $ka$ and $\bar{k}\bar{a}$ are indistinguishable, there is only one distinguishable path $l$ associated with them. We use function $e(\cdot)$ to map a given path into a distinguishable path, for example, $e(ka) = e(\bar{k}\bar{a}) = l$.

Let $y_l \in \{0,1\}$, for $l \in \mathscr{L}_o, o \in \mathscr{B}$. Then $y_{e(ka)} = 1$ if commodity $k$ is allowed to use hub arc $a$ (the same applies for commodity $\bar{k}$ and hub arc $\bar{a}$ if the path $\bar{k}\bar{a}$ does exist); otherwise $y_{e(ka)} = 0$. Constraints (18) are equivalent to

$$\lambda_{a_1} \leq f^{ka}(\lambda_{a_2}), \qquad \text{if } y_{e(ka)} = 1, k \in K, a \in \mathscr{A}_k, \tag{21}$$

$$x_{ka} \leq y_{e(ka)}, \qquad \forall k \in K, a \in \mathscr{A}_k. \tag{22}$$

The function $f^{ka}$ is approximated by a set of tangent lines at breakpoints $\lambda^r_{a_2}$, for $r \in R$, as shown in Figure 3. Assuming the convexity of probabilistic service level constraints, $f^{ka}$ is concave, and its piece-wise linear approximation is given by:

$$f^{ka}(\lambda_{a_2}) \approx \min_{r \in R} \left\{ f^{ka}(\lambda^r_{a_2}) + \left(\lambda_{a_2} - \lambda^r_{a_2}\right) f^{ka}_1(\lambda^r_{a_2}) \right\}, \qquad \forall r \in R,$$

which is a strictly decreasing outer approximation of $f^{ka}$. Then, an approximation of probabilistic service level (21) constraints is obtained by the following set of disjunctive linear inequalities:

$$\lambda_{a_1} \leq f^{ka}(\lambda^r_{a_2}) + \left(\lambda_{a_2} - \lambda^r_{a_2}\right) f^{ka}_1(\lambda^r_{a_2}), \qquad \text{if } y_{e(ka)} = 1, k \in K, a \in \mathscr{A}_k, r \in R. \tag{23}$$

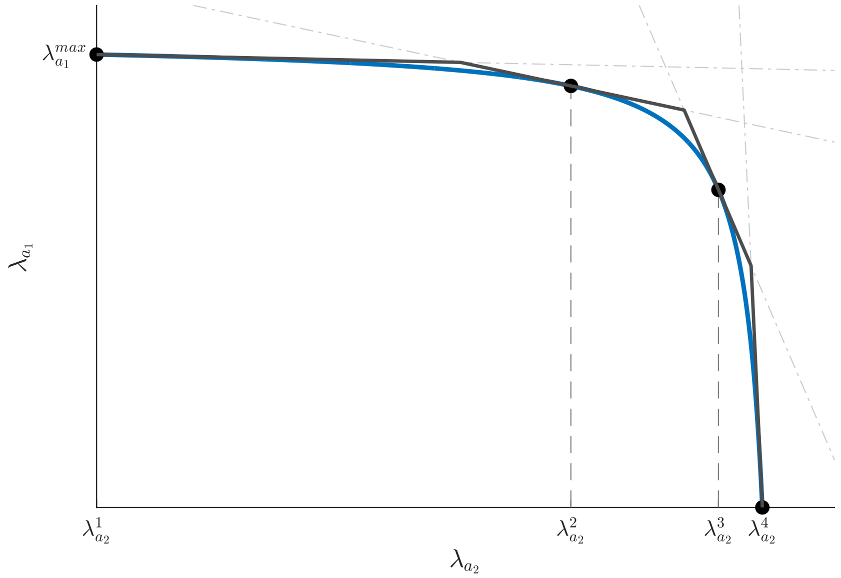


**Figure 3** **Piece-wise linear approximation of function $f^{ka}$ using four line segments with tangent points $\{(\lambda^r_{a_2}, f^{ka}(\lambda^r_{a_2}))\}$, for $r \in R = \{1,\ldots,4\}$.**

### 5.3. Perspective cuts

From the logical condition in Constraints (23), we have two sets that represent feasible regions for the pair $(\lambda_{a_1}, \lambda_{a_2})$, depending on the value of the binary variable $y_{e(ka)}$:

$$R^{0a} = \left\{(\lambda_{a_1}, \lambda_{a_2}) \in \mathbb{R}^2 \mid \lambda_{a_1} \in [0, \Lambda_{a_1}], \lambda_{a_2} \in [0, \Lambda_{a_2}]\right\}, \tag{24}$$

$$R^{ka} = \left\{(\lambda_{a_1}, \lambda_{a_2}) \in \mathbb{R}^2 \mid \lambda_{a_1} \leq f^{ka}(\lambda_{a_2}), \lambda_{a_1} \geq 0, \lambda_{a_2} \geq 0\right\}. \tag{25}$$

Assuming the validity of Conjecture 1, the set $R^{ka}$ is convex and defines the feasible set for $\lambda_{a_1}$ and $\lambda_{a_2}$ when $y_{e(ka)} = 1$. For its part, the set $R^{0a}$ is convex and it is considered as an alternative when $y_{e(ka)} = 0$. It is of interest to produce a formulation whose relaxation coincides with the convex hull of set $R^{0a} \cup R^{ka}$, which is denoted as $\mathbf{conv}(R^{0a} \cup R^{ka})$ (see Figure 4a). Given that such a formulation is nonlinear, in this paper, we approximate $\mathbf{conv}(R^{0a} \cup R^{ka})$ using linear inequalities.

Proposition 5 **(Fine perspective cuts).** *Let $\lambda^r_{a_2}$ be a valid breakpoint for $f^{ka}$ and assume the validity of Conjecture 1. The following inequalities are valid and represent the strongest linear cuts for $\boldsymbol{conv}(R^{0a} \cup R^{ka})$:*

$$\lambda_{a_1} + \left|f^{ka}_1(\lambda^r_{a_2})\right| \lambda_{a_2} \leq \left[f^{ka}(\lambda^r_{a_2}) + \lambda^r_{a_2}\left|f^{ka}_1(\lambda^r_{a_2})\right|\right] y_{e(ka)} + \left[\Lambda_{a_1} + \left|f^{ka}_1(\lambda^r_{a_2})\right| \Lambda_{a_2}\right](1 - y_{e(ka)}), \tag{26}$$

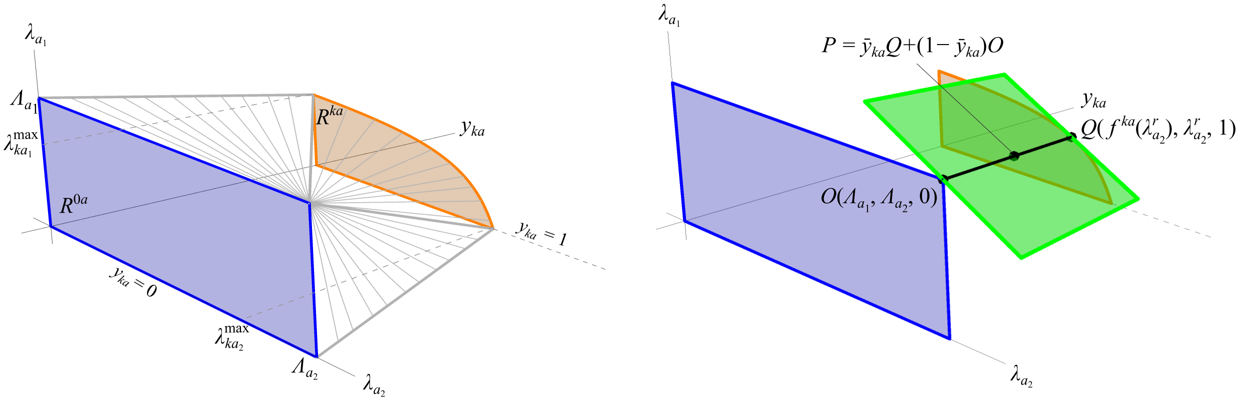
$$\lambda_{a_1} \leq \lambda^{max}_{ka_1}(a_2) y_{e(ka)} + \Lambda_{a_1}(1 - y_{e(ka)}), \tag{27}$$

$$\lambda_{a_2} \leq \lambda^{max}_{ka_2}(a_1) y_{e(ka)} + \Lambda_{a_2}(1 - y_{e(ka)}), \tag{28}$$

*where $|\cdot|$ is the absolute value function.*

*Proof* We present a proof for Inequality (26). The proof for the other inequalities holds similarly. Consider points $O = (\Lambda_{a_1}, \Lambda_{a_2}, 0) \in R^{0a}$ and $Q = \left(f^{ka}(\lambda^r_{a_1}), \lambda^r_{a_2}, 1\right) \in R^{ka}$, let $\bar{y}_{ka} \in [0,1]$, and $P = \bar{y}_{ka} Q + (1 - \bar{y}_{ka}) O$, hence $P \in \mathbf{conv}(R^{0a} \cup R^{ka})$. Let $\Gamma$ be the set of points that satisfy Inequality (26) as strict equality. Evaluating $P$ in Inequality (26) produces a strict equality implying that $P \in \Gamma$. Now hyperplane $\Gamma$ is tangent to $R^{0a}$ and $R^{ka}$ at points $Q$ and $O$, respectively, and given that $R^{0a}$ and $R^{ka}$ are convex, we conclude that inequality (29) is valid and is the strongest possible inequality for $\mathbf{conv}(R^{0a} \cup R^{ka})$ at break point $\lambda^r_{a_2}$. □

Figure 4 illustrates Proposition 5. We refer to constraints of the form (29) as *perspective cuts*. Frangioni and Gentile (2006) introduced these cuts for approximating the convex hull of the union of a convex set and a point. We use perspective cuts for the convex hull of the union of a convex set and a box. The formulation of the convex hull using non-linear constraints is not considered in this paper, but the interested reader is referred to Hijazi et al. (2012).



(a) Convex hull of the set $R^a \cup R^{ka}$.

(b) Perspective cut.

**Figure 4** **Nature of perspective cuts.**

Inequality (26) can be rewritten considering the impact of location variables $z_m$, as shown below. This is valid since the left-hand side is non-negative and, by Constraints (2c) and (2d), $\lambda_{a_1}$ ($\lambda_{a_2}$) is zero when $z_{a_1}$ ($z_{a_2}$) is zero, and $y_{e(ka)}$ is zero when $z_{a_1}$ or $z_{a_2}$ is zero.

$$\lambda_{a_1} + \left|f_1^{ka}(\lambda_{a_2}^r)\right| \lambda_{a_2} + \left[\Lambda_{a_1} - f^{ka}(\lambda_{a_2}^r) + \left(\Lambda_{a_2} - \lambda_{a_2}^r\right)\left|f_1^{ka}(\lambda_{a_2}^r)\right|\right] y_{e(ka)} \leq \Lambda_{a_1} z_{a_1} + \left|f_1^{ka}(\lambda_{a_2}^r)\right| \Lambda_{a_2} z_{a_2}$$
$$\forall k \in K, a \in \mathscr{A}_k, r \in R. \qquad (29)$$

Similarly, Constraints (27) and(30) are written as

$$\left[\Lambda_{a_1} - \lambda_{ka_1}^{\max}(a_2)\right] y_{e(ka)} + \lambda_{a_1} \leq \Lambda_{a_1} z_{a_1}, \qquad \forall k \in K, a \in \mathscr{A}_k, a_1 \in a. \qquad (30)$$

It is important to note that Proposition 5 establishes the strongest possible linear cuts for $\mathbf{conv}(R^{0a} \cup R^{ka})$. However, it does not imply that these cuts are the strongest possible for the entire optimization problem because the problem contains similar disjunctive constraints for other hub arcs, which are considered in an isolated way. Another issue is that there is a large number of inequalities (29), and it is inefficient to include all of them at the root node of the branch-and-bound algorithm.

### 5.4. Cutting plane algorithm

With the new definitions of sets, parameters, and valid inequalities, Formulation (2) is modified as follows.

$$(M1): \text{ minimize } \quad \sum_{m \in \mathcal{N}} f_m z_m - \sum_{k \in K} \sum_{a \in \mathscr{A}_k} (C_k - C_{ka}) x_{ka} + \sum_{k \in K} C_k \qquad (31a)$$

$$\text{subject to} \quad \sum_{a \in \mathscr{A}_k} x_{ka} \leq 1, \qquad \forall k \in K, \qquad (31b)$$

$$\sum_{\substack{a \in \mathscr{A}_k, \\ m \in a}} x_{ka} \leq z_m, \qquad \forall m \in \mathcal{N}, k \in K, \qquad (31c)$$

$$\sum_{k\in K}\sum_{\substack{a\in\mathscr{A}_k,\\ m\in a}} w_k x_{ka} \le \Lambda_m z_m, \qquad \forall m\in\mathscr{N}, \tag{31d}$$

$$x_{ka} \le y_{e(ka)}, \qquad \forall k\in K, a\in\mathscr{A}_k \tag{31e}$$

$$(29), (30)$$

$$x_{ka}\in[0,1], \qquad \forall k\in K, a\in\mathscr{A}_k \tag{31f}$$

$$z_m\in\{0,1\}, \qquad \forall m\in\mathscr{N}, \tag{31g}$$

$$y_l\in\{0,1\}, \qquad \forall l\in\mathscr{L}_o, o\in\mathscr{B}. \tag{31h}$$

We do not include cuts (29) and (30) in the root-node relaxation. Those cuts are added as cutting planes during the branch-and-bound process when it is necessary to produce feasible solutions. These *feasibility cuts* are added only at integer nodes of the branching tree.

Let $\varepsilon$ be the tolerance error for probabilistic service level constraints. Let $(\bar{\mathbf{x}},\bar{\mathbf{y}},\bar{\mathbf{z}})$ be a feasible solution of a node of the branching tree, where $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ are integer. The procedure for generating cutting planes at integer nodes of the branching tree is described in Algorithm 1.

**Algorithm 1:** Generating cutting planes at integer nodes of the branching tree of $(M1)$

```
Data: $\mathscr{B}, \mathscr{L}_o, \varepsilon, \alpha, (\bar{\mathbf{x}},\bar{\mathbf{y}},\bar{\mathbf{z}})$, parameters related to the total service time distribution.
for $o=(m,n)\in\mathscr{B}$ do
    $CutCnt \leftarrow 0$;
    if $\bar{z}_m = 1$ and $\bar{z}_n = 1$ then
        $\lambda_m \leftarrow \sum_{k\in K}\sum_{a\in\mathscr{A}_k, m\in a} w_k\bar{x}_{ka}$;
        $\lambda_n \leftarrow \sum_{k\in K}\sum_{a\in\mathscr{A}_k, n\in a} w_k\bar{x}_{ka}$;
        for $l\in\mathscr{L}_o$ do
            if $\bar{y}_l = 1$ then
                if $\lambda_n > \lambda^{max}_{ln}(m)$ or $\lambda_m > \lambda^{max}_{lm}(n)$ then
                    Add cuts (30);
                else if $S_{T_{lo}}(\tau_l \mid \lambda_m, \lambda_n) < \alpha - \varepsilon$ then
                    Add cut (29);
                end
            end
        end
    end
end
```

## 6. Approximations and valid inequalities

In this section, we present an approximated formulation for probabilistic service level constraints using a set of homothetic functions to approximate the $\alpha$-level sets of the total service time distribution for the different commodities. We then propose an extended formulation for the exact problem.

### 6.1. Homothetic outer approximations

We propose to approximate functions $f^{ka}$, for $k \in K, a \in \mathscr{A}_k$, using homothetic functions $\tilde{f}^{ka}$ with common homothetic center. Function $f^{ka}$ defines the $\alpha$-level set of the total sojourn time distribution $S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2})$ (see Section 4). For its part, $S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2})$ is related to the total sojourn time distribution $W_{V_a}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2})$ by means of the Convolution Integral (12). From this integral, $S_{T_{ka}}$ can be interpreted as a sort of "incomplete" expectation of functions $W_{V_a}$ defined for parameter $\tau \in [0, \tau_k]$, where the expectation is taken considering some affine transformation of the density function $g_{U_{ka}}$. Because of the convolution operation, it is not surprising that $S_{T_{ka}}$ and $W_{V_a}$ share many similarities.

Intending to exploit the homothetic properties of the total sojourn time distribution (see Proposition 4), we propose the approximation $S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2}) \approx W_{V_a}(\tilde{\tau}_{ka} \mid \lambda_{a_1}, \lambda_{a_2})$, for some $\tilde{\tau}_{ka}$. Let $\tilde{f}^{ka}$ be the function that defines the $\alpha$-level set of $W_{V_a}(\tilde{\tau}_{ka} \mid \lambda_{a_1}, \lambda_{a_2})$. It is convenient to define parameter $\tilde{\tau}_{ka}$ such that $\tilde{f}^{ka} \geq f^{ka}$. By the monotonicity properties of $W_{V_a}$, there always exists $\tilde{\tau}_{ka} \in [0, \tau_k]$ for which such inequality (outer approximation) is met. To see this, note that given that $W_{V_a}$ is strictly increasing to time we have

$$\begin{aligned} S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2}) = &\int_0^{\tau_k} W_{V_a}(\tau_k - x \mid \lambda_{a_1}, \lambda_{a_2}) g_{U_{ka}}(x) dx < \int_0^{\tau_k} W_{V_a}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2}) g_{U_{ka}}(x) dx, \\ &< W_{V_a}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2}). \end{aligned}$$

Given that the total service time and the total sojourn time distributions are strictly decreasing and continuous to $\lambda_{a_1}$ and $\lambda_{a_2}$, we have that $C_\alpha^{S_{T_{ka}}(\tau_k)} \subseteq C_\alpha^{W_{V_a}(\tau_k)}$, see Proposition 6 below for an analogous reasoning. As a consequence, $\tilde{f}^{ka} > f^{ka}$, for $\tilde{\tau}_{ka} = \tau_k$. Algorithm 2 in Appendix F is a heuristic suggested for determining a convenient value for $\tilde{\tau}_{ka}$.

Proposition 6. *Let $a = (a_1, a_2)$ and $k, l \in \mathscr{K}_{a_1} \cap \mathscr{K}_{a_2}$. If $\tilde{\tau}_{ka} \leq \tilde{\tau}_{la}$, then $\tilde{f}^{ka} \leq \tilde{f}^{la}$, equivalently, $\tilde{R}^{la} \supseteq \tilde{R}^{ka}$.*

*Proof* Given that $W_{V_a}$ is increasing with respect to time, we have $W_{V_a}(\tilde{\tau}_{ka} \mid \lambda_{a_1}, \lambda_{a_2}) \leq W_{V_a}(\tilde{\tau}_{la} \mid \lambda_{a_1}, \lambda_{a_2})$. Now, given that $W_{V_a}$ is strictly decreasing and continuous with respect to $\lambda_{a_1}$ and $\lambda_{a_2}$, there exist $\epsilon \geq 0$ such that $\alpha = W_{V_a}(\tilde{\tau}_{ka} \mid \lambda_{a_1}, \lambda_{a_2}) = W_{V_a}(\tilde{\tau}_{la} \mid \lambda_{a_1} + \epsilon, \lambda_{a_2})$, implying that $\lambda_{a_1} = \tilde{f}^{ka}(\lambda_{a_2}) \leq \lambda_{a_1} + \epsilon = \tilde{f}^{la}(\lambda_{a_2})$. □

Definition 2 (Homothetic ordering). The homothetic ordering of commodities concerning hub arc $a$ is the list of commodities ordered such that

$$\tilde{R}^{[1]a} \supseteq \tilde{R}^{[2]a} \supseteq \cdots \supseteq \tilde{R}^{[k]a} \supseteq \cdots \supseteq \tilde{R}^{[|\mathscr{K}_{a_1} \cap \mathscr{K}_{a_2}|]a}, \tag{32}$$

where $[k]$ refers to the commodity at the $k$-th position of the list, and $a \in \mathscr{B}$.

### 6.2. Incremental formulation for the approximated problem

Let

$$\tilde{\lambda}^{\max}_{ka_1}(a_2) = \tilde{f}^{ka}(0), \qquad \forall k \in K, a = (a_1, a_2) \in \mathscr{A}_k,$$
$$L_m = \max\left\{\Lambda_m, \max_{k\in K, n\in\mathscr{N}} \left\{\tilde{\lambda}^{\max}_{km}(n)\right\}\right\}, \qquad \forall m \in \mathscr{N}.$$

Figure 5 illustrates the family of functions $\left\{\tilde{f}^{ka}\right\}$ for hub arc $a$. A ray $\mathbf{r}$ from the homothetic center $(\mu_{a_1}, \mu_{a_2})$ cuts each curve at points $(\lambda^{\mathbf{r}}_{k_ia_1}, \lambda^{\mathbf{r}}_{k_ia_2})$, for $i = 1, 2, \ldots$, with the same derivative $s^{\mathbf{r}}_a$. In this approximation, a valid upper bound for $\lambda_m$ is $L_m$. This upper bound is valid irrespective of which hubs are open. The bounds imposed to flow variables imply that rays $\mathbf{r}$ are also bounded. To formalize the definition of this bound, we represent a ray as a vector with origin at the point $(\mu_{a_1}, \mu_{a_2})$. Let $\bar{\mathbf{r}}_a = (L_{a_2} - \mu_{a_2}, -\mu_{a_1})$ and $\underline{\mathbf{r}}_a = (-\mu_{a_2}, L_{a_1} - \mu_{a_1})$ be the vectors representing the rays that contain the points $(0, L_{a_1})$ and $(L_{a_2}, 0)$, respectively, as shown in Figure 5.

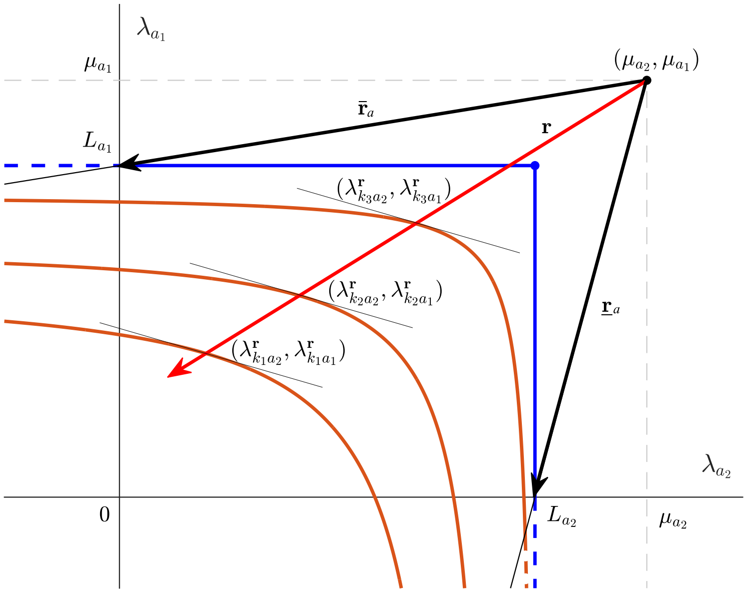

$\lambda_{a_1}$
$\mu_{a_1}$
$(\mu_{a_2}, \mu_{a_1})$
$\bar{\mathbf{r}}_a$
$\mathbf{r}$
$L_{a_1}$
$(\lambda^{\mathbf{r}}_{k_3a_2}, \lambda^{\mathbf{r}}_{k_3a_1})$
$(\lambda^{\mathbf{r}}_{k_2a_2}, \lambda^{\mathbf{r}}_{k_2a_1})$
$\underline{\mathbf{r}}_a$
$(\lambda^{\mathbf{r}}_{k_1a_2}, \lambda^{\mathbf{r}}_{k_1a_1})$
$\lambda_{a_2}$
0
$L_{a_2}$
$\mu_{a_2}$

**Figure 5 Geometric construction considering the homothetic outer approximations.**

Definition 3 (Relevant rays). We say that ray $\mathbf{r}$ is relevant for hub arc $a$ if $\mathbf{r} \in \mathscr{C}_a = \{c_1\underline{\mathbf{r}}_a + c_2\bar{\mathbf{r}}_a \mid c_1, c_2 \geq 0\}$.

From the geometric construction in Figure 5 and Definition 3, we note that a relevant ray $\mathbf{r}$ intersect function $\tilde{f}^{ka}$ at a point

$$(\lambda^{\mathbf{r}}_{ka_1}, \lambda^{\mathbf{r}}_{ka_2}) \in \{(\lambda_{a_1}, \lambda_{a_2}) \mid \lambda_{a_1} \in (-\infty, L_{a_1}], \lambda_{a_2} \in (-\infty, L_{a_2}]\} \cap \mathscr{C}_a. \tag{33}$$

Let $v_{e(ka)} \in \{0,1\}$ be an indicator variable with result $v_{e(ka)} = 1$ if commodity $k$ is allowed to use hub arc $a$ in the approximated problem; otherwise $v_{e(ka)} = 0$. By considering the homothetic ordering of commodities, variables $v_{e(ka)}$ can be interpreted as incremental variables.

PROPOSITION 7. *Consider the list of transport paths $\mathscr{L}_a$ for hub arc $a \in \mathscr{B}$ ordered following Definition 2 for strictly distinguishable commodities. Let $[l] \in \mathscr{L}_a$ be the $l$-th element of the list. The inequalities: $v_{[l]} \le v_{[l-1]}$, $v_{[1]} \le z_{a_1}$, and $v_{[1]} \le z_{a_2}$, for $[1] \in \mathscr{L}_a$, are valid.*

*Proof* Consider the proof for the incremental inequality $v_{[l]} \le v_{[l-1]}$ first. Let transport path $[l]$ be associated with commodity $k$, and transport path $[l-1]$ be associated with commodity $k'$. Assume that commodity $k$ is allowed to use hub arc $a$, i.e., $v_{[l]} = 1$. By Proposition 6 we have that $\tilde{f}^{ka}(\lambda_{a_2}) \le \tilde{f}^{k'a}(\lambda_{a_2})$, implying that commodity $k'$ is also allowed to use hub arc $a$.

Now consider inequalities $v_{[1]} \le z_{a_1}$ and $v_{[1]} \le z_{a_2}$. If either hub $a_1$ or hub $a_2$ is closed, then $v_{[1]} = 0$ and $v_{[1]} = 0$, by the incremental inequalities, we also have that $v_l = 0$ for all $l \in \mathscr{L}_a$. This implies that no commodity is allowed to use hub arc $(a_1, a_2)$, as it is required when at least one of the hubs at the ends of the arc is closed. □

PROPOSITION 8 **(Coarse perspective cut)**. *Let $\mathbf{r} \in \mathscr{C}_a$, $(\lambda^{\mathbf{r}}_{ka_1}, \lambda^{\mathbf{r}}_{ka_2})$ be the intersection point of ray $\mathbf{r}$ and function $\tilde{f}^{ka}$, $s^{\mathbf{r}}_a$ be the common derivative or slope of the tangent lines at intersection points, and consider that $v_{[l]} \le v_{[l-1]}$, for $[l] \in \mathcal{L}_a$. The following inequality is valid for the approximated version of probabilistic service level constraints:*

$$\lambda_{a_1} + |s^{\mathbf{r}}_a|\lambda_{a_2} + \sum_{l \in \mathscr{L}_a} \tilde{\Delta}^{\mathbf{r}}_l v_l \le L_{a_1} z_{a_1} + |s^{\mathbf{r}}_a| L_{a_2} z_{a_2}, \qquad \forall a \in \mathscr{B}, \mathbf{r} \in \mathscr{C}_a. \tag{34}$$

*where*

$$\tilde{\Delta}^{\mathbf{r}}_{[l]} = \left\{ \begin{matrix} \tilde{\delta}^{\mathbf{r}}_{[l]} - \tilde{\delta}^{\mathbf{r}}_{[l-1]}, & \textit{for } l > 1, \\ \tilde{\delta}^{\mathbf{r}}_{[l]}, & \textit{for } l = 1 \end{matrix} \right\} \ge 0, \qquad \forall [l] \in \mathcal{L}_a,$$

*and* $\tilde{\delta}^{\mathbf{r}}_l = \tilde{\delta}^{\mathbf{r}}_{e(ka)} = L_{a_1} - \lambda^{\mathbf{r}}_{ka_1} + |s^{\mathbf{r}}_a| \left(L_{a_2} - \lambda^{\mathbf{r}}_{ka_2}\right)$.

*Proof* See Appendix G. □

An important difference between the *coarse perspective cut* (34) compared to the *fine perspective cut* (29) is that the former does not need to be defined for each commodity. This reduces the number of constraints required considerably. The formulation for the approximated problem is as follows:

$$(M2): \text{ minimize } \quad \sum_{m \in \mathcal{N}} f_m z_m - \sum_{k \in K} \sum_{a \in \mathscr{A}_k} (C_k - C_{ka}) x_{ka} + \sum_{k \in K} C_k \tag{35a}$$

$$\text{subject to} \qquad (31b) - (31d), (31f), (31g), (34)$$

$$x_{ka} \le v_{e(ka)}, \qquad \forall k \in K, a \in \mathscr{A}_k, \tag{35b}$$

$$v_{[l]} \leq v_{[l-1]}, \qquad \forall a \in \mathscr{B}, [l] \in \mathscr{L}_a, \tag{35c}$$

$$v_{[1]} \leq z_m, \qquad \forall a \in \mathscr{B}, [1] \in \mathscr{L}_a, m \in a, \tag{35d}$$

$$v_l \in \{0,1\}, \qquad \forall a \in \mathscr{B}, l \in \mathscr{L}_a. \tag{35e}$$

We do not include any Coarse perspective cuts (34) in the root relaxation. These cuts are added if the approximated probabilistic service level constraints are not satisfied up to a tolerance error $\tilde{\varepsilon}$. Omitting lines 20–22 from Algorithm 3 in Appendix J, this algorithm is used to solve the approximated formulation. Lines 20–22 are reserved only for the solution of model $(M3)$, as discussed below.

There is an alternative formulation for the approximated problem. We refer to this formulation as the *multiple choice* formulation and it is described in Appendix H. According to our experience, the incremental formulation outperforms the multiple-choice formulation in terms of processing time.

### 6.3. Introducing finer cuts

The effectiveness of the coarse perspective cuts in the approximated formulation relies on the reduced number of cuts required for each hub arc and in the incremental nature of binary variables $v_l$, which is a consequence of the homothetic ordering of commodities (see Proposition 7). Unfortunately, we cannot establish a similar ordering for the exact version of probabilistic service level constraints. In this case, we might have some commodities that violate such an ordering at some points $(\lambda_{a_1}, \lambda_{a_2}) \in D$. Consequently, Proposition 7 is no longer valid for the exact formulation $(M1)$. Appendix I presents some valid inequalities for formulation $(M1)$, but considering such inequalities does not seem to improve the computational performance.

A different approach is to consider an extended formulation with two sets of binary variables, the binary variables $y_l$ used to model fine perspective cuts, along with the binary variables $v_l$ used to model coarse perspective cuts. We use a cutting plane algorithm by introducing coarse and fine perspective cuts. The idea is to introduce coarse perspective cuts (34) at integer solutions of the branching tree with a tolerance error $\tilde{\varepsilon}$. Fine perspective cuts (29) are introduced with a tolerance error of $\varepsilon < \tilde{\varepsilon}$ when the approximated version of probabilistic service level constraints is satisfied for all commodities and hub arcs in the current integer node of the branching three. We start with Formulation (36):

$$(M3): \text{ minimize } \quad \sum_{m \in \mathscr{N}} f_m z_m - \sum_{k \in K} \sum_{a \in \mathscr{A}_k} (C_k - C_{ka}) x_{ka} + \sum_{k \in K} C_k \tag{36a}$$

$$\text{subject to} \quad (31b) - (31e), (31f) - (31h), (35b) - (35e), (34).$$

We do not include constraints $x_{ka} \leq y_{e(ka)}$ and any fine or coarse perspective cut in the root relaxation. Previous constraints are added when necessary during the branch-and-bound algorithm.

To solve formulation $(M3)$, we use Algorithm 3 including lines 22–23. The procedure is described in Algorithm 4 in Appendix J.

### 6.4. Numerical validation of the convexity of probabilistic service level constraints

It has been shown that probabilistic service level constraints are of the form $\lambda_1 \leq f(\lambda_2)$. Assuming the validity of Conjecture 1, these constraints are convex, implying that $f$ is concave. We propose the linear outer approximation: $\lambda_1 \leq \min_{r \in R} \{f(\lambda_2^r) + (\lambda_2 - \lambda_2^r) f_1(\lambda_2^r)\}$, where $\lambda_2^r$ are sample points in the domain of $f$. This linear outer approximation is concave if, for a strictly increasing sequence $\{\lambda_2^1, \lambda_2^2, \ldots \lambda_2^{|R|}\}$, the associated slope sequence $\{f_1(\lambda_2^1), f_1(\lambda_2^2), \ldots f_1(\lambda_2^{|R|})\}$ is strictly decreasing. We also check the convexity of the homothetic outer approximation $\lambda_1 \leq \tilde{f}(\lambda_2)$. In this case, the sequence $\{\mathbf{r}\}$ corresponds to the list of rays ordered in a counterclockwise manner. The corresponding sequence of slopes $\{s_a^{\mathbf{r}}\}$ must follow a strictly decreasing sequence for validating convexity. Convexity was verified in this way for all instances of our numerical experiments.

## 7. Computational experiments

We conducted an extensive computational study to assess the performance of the three formulations over problem instances under different service levels. The proposed formulations are coded in C and executed on a Dell PowerEdge R740 PC, 2 Intel Xeon Gold 6258R, CPU running at 2.70 GHz, and 60 GB RAM. The formulations are solved using the CPLEX 22.1.1 Callable Library with its default settings using one thread. We utilize the GNU Scientific Library for the required numerical computations.

### 7.1. Test Instances

We performed our experiments on a total of 480 instances generated from two datasets - the Australian Post (AP) dataset (Ernst and Krishnamoorthy 1996) and the Colombian (COL) dataset Basallo-Triana et al. (2023). We considered 240 problem instances for each dataset. The AP dataset consists of postal flow and Euclidean distances between 200 districts in an Australian city and can be downloaded from the OR library. In the AP dataset, the capacities and fixed costs are defined in two distinct scenarios: loose (L) and tight (T), respectively. For the AP dataset, we set the unit transportation costs as follows: $\chi = 3$, $\tau = 0.75$, $\delta = 2$, $\rho = 7$. Additionally, the hub installation costs $f_m$ from the original dataset are multiplied by a factor of 0.5.

The Colombian (COL) dataset, introduced by Basallo-Triana et al. (2023), contains the fixed installation costs of hubs for various capacity levels. The processing rate of the flow that arrives at a hub from non-hub nodes (export flow) differs from the processing rate of the flow that arrives at a hub from other hubs (import flow). Given that here we are assuming that import and export flows are processed at the same rate, we use the average processing rate (of import and export units) to define the hub capacity. In all our experiments, we have considered the second capacity level of the COL dataset to define the hub capacity. In this case, we set the unit transportation cost as follows: $\chi = 1$, $\tau = 0.75$, $\delta = 1$, $\rho = 1.5$. The dataset is available at https://github.com/MarioBasallo/COL_dataset.git.

The maximum service time requirement $\tau_k$ is defined as a factor $r$ of the $100\ell$-th percentile of the direct transportation time density function of commodity $k$. We assume that the direct transport time of commodity $k$ has a gamma density function with shape parameter $a_k = 1/CV^2$ and scale parameter $b_k = (c_k/v)(1/a_k)$, where $CV$ is the coefficient of variation of the density function, $c_k$ is the direct transportation distance, and $v$ is the travel speed of a vehicle. The travel speed is computed as $v = 0.2\bar{\mu}\bar{d}$, where $\bar{\mu}$ and $\bar{d}$ are the average processing rate and the average direct transportation distance, respectively. The maximum service time requirement is computed as follows:

$$\tau_k = rG_k^{-1}(\ell \mid a_k, b_k), \tag{37}$$

where $G_k^{-1}$ is the inverse of the cumulative gamma distribution function for the direct transportation time of commodity $k$. We set $\ell = 0.7$ and consider that $r > 1$, which is motivated by a situation where direct transport is faster than combined rail-road transport. A small value of $r$ leads to a tighter service time requirement $\tau_k$ for combined rail-road transportation.

To compute the intermodal transportation time, we use a gamma density function with parameters $a_{ka} = 1/CV^2$ and $b_{ka} = E[U_{ka}]/a_{ka}$, where $E[U_{ka}]$ is the expected total transportation time for commodity $k$ using hub arc $a$, which is computed as follows:

$$E[U_{ka}] = E[U_{(i,j)(a_1,a_2)}] = \frac{1}{v}\left(c_{ia_1} + \eta c_{a_1a_2} + c_{a_2j}\right). \tag{38}$$

In this case, $\eta$ is a time correction factor used to account for the differences in travel speeds between trucks and trains. We set $\eta = 1.5$, implying that trains (inter-hub vehicles) are slower than trucks (hub-and-spoke vehicles).

Service levels are selected from the set $\alpha = \{0.80, 0.85, 0.90, 0.95, 0.99\}$. The factor $r$ varies within the set $r \in \{2, 3, 4\}$. In the AP dataset, the instances are labeled as $n$FC, where $n \in \{10, 20, 25, 40\}$ and FC $\in$ {ll, lt, tl, tt}. Note that the AP dataset does not include networks with $n = 30$ nodes, hence we do not consider $n = 30$ in our experiments. For the COL dataset, the instances are labeled as $n$-$m$, where $m \in \{1, 2, 3, 4\}$, and $n \in \{10, 20, 25, 40\}$. The tolerance error for the probabilistic service level constraints is set to $\varepsilon = 10^{-6}$. We conducted experiments with three formulations: $(M1)$ (31), $(M2)$ (35) and $(M3)$ (36). For $(M2)$, we set the error to $\tilde{\varepsilon} = 10^{-5}$, and for $(M3)$, we set the error to $\tilde{\varepsilon} = 10^{-1}$. In Formulation $(M1)$, we add fine perspective cuts during the branch-and-bound algorithm. Formulation $(M2)$ considers the addition of coarse perspective cuts exclusively. In formulation $(M3)$, we add both fine and coarse perspective cuts. The source code for the algorithms is available at https://github.com/MarioBasallo/IHND.git.

### 7.2. Accuracy of the homothetic approximations

Appendix K shows an analysis of the accuracy of Algorithm 2 for obtaining homothetic approximations. The average absolute error for approximating the service level is of the order of $10^{-3}$, which demonstrates the effectiveness of the algorithm by providing highly accurate approximations.

We compare the optimal solutions of the approximated formulation ($M2$) with that of the exact formulation ($M1$). The results are reported in Tables 2 and 3. We note that the intermodal hub networks produced by the approximate formulation differ slightly from the exact formulation in terms of the minimum service level guarantee, but the difference is rather small (less than 1.0%). Furthermore, only a small fraction of commodities violate probabilistic service level constraints. For its part, the average gap between the objective function values of the optimal solution and the solution of the approximate formulation is also small (less than 0.9%). The fact that the gap is non-negative and that probabilistic service level constraints are violated slightly in the approximate formulation reinforces the idea that Algorithm 2 produces an accurate outer approximation for the non-linear probabilistic service level constraints.

**Table 2 Summary of results for the approximated and the exact model for the AP dataset.**

| | | Minimum network service level (%) | | Gap* (%) | |
|---|---|---|---|---|---|
| $r$ | $\alpha$ (%) | ($M1$) | ($M2$)† | Minimum | Maximum |
| 2 | 80 | 80.000 | 79.945 (98.3) | 0.000 | 0.064 |
| | 85 | 85.000 | 84.969 (98.2) | 0.000 | 0.311 |
| | 90 | 90.000 | 89.942 (98.7) | 0.000 | 0.111 |
| | 95 | 95.022 | 95.022 (100) | 0.000 | 0.000 |
| | 99 | NH** | NH** | 0.000 | 0.000 |
| 3 | 80 | 80.000 | 79.810 (99.1) | 0.000 | 0.049 |
| | 85 | 85.000 | 84.905 (98.5) | 0.000 | 0.043 |
| | 90 | 90.000 | 89.864 (98.6) | 0.000 | 0.105 |
| | 95 | 95.000 | 94.836 (97.8) | 0.000 | 0.369 |
| | 99 | 99.000 | 98.935 (98.2) | 0.000 | 0.593 |
| 4 | 80 | 80.000 | 79.766 (99.2) | 0.000 | 0.031 |
| | 85 | 85.000 | 84.787 (98.5) | 0.000 | 0.055 |
| | 90 | 90.000 | 89.709 (98.2) | 0.000 | 0.229 |
| | 95 | 95.000 | 94.730 (98.2) | 0.000 | 0.351 |
| | 99 | 99.000 | 98.894 (96.3) | 0.009 | 0.809 |

* Gap: $100(\texttt{M1} - \texttt{M2})/\texttt{M1}$.
** No hubs open.
† The numbers in parenthesis refer to the percentage of commodities that satisfy probabilistic service level constraints.

**Table 3** Summary of results of the approximated and the exact model for the COL dataset.

| | | Minimum network service level (%) | | Gap (%) | |
|---|---|---|---|---|---|
| $r$ | $\alpha$ (%) | $(M1)$ | $(M2)$ | Minimum | Maximum |
| 2 | 80 | 80.000 | 79.900 (93.2) | 0.000 | 0.042 |
| | 85 | 85.000 | 84.915 (96.5) | 0.000 | 0.029 |
| | 90 | 90.000 | 89.893 (97.7) | 0.000 | 0.159 |
| | 95 | NH | 94.970 (98.7) | 0.000 | 0.008 |
| | 99 | NH | NH | 0.000 | 0.000 |
| 3 | 80 | 80.000 | 79.768 (94.7) | 0.002 | 0.038 |
| | 85 | 85.000 | 84.872 (94.3) | 0.003 | 0.029 |
| | 90 | 90.000 | 89.872 (94.8) | 0.007 | 0.076 |
| | 95 | 95.000 | 94.773 (92.4) | 0.023 | 0.217 |
| | 99 | 99.000 | 98.915 (93.2) | 0.000 | 0.343 |
| 4 | 80 | 80.000 | 79.815 (95.2) | 0.004 | 0.127 |
| | 85 | 85.000 | 84.794 (95.0) | 0.004 | 0.054 |
| | 90 | 90.000 | 89.767 (94.8) | 0.010 | 0.088 |
| | 95 | 95.000 | 94.770 (94.8) | 0.000 | 0.204 |
| | 99 | 99.000 | 98.878 (90.6) | 0.048 | 0.601 |

### 7.3. Computational performance

Tables 4 and 5 show the descriptive statistics for the numerical results, for the AP and COL datasets, respectively. The columns labeled *Min*, *Avg*, *Max*, and *St. Dev.* refer to the minimum, average, maximum, and standard deviation of the computational time (in seconds) of the instances solved to optimality. The column labeled *Cut time* shows the average time of the execution of the lazy constraint function in CPLEX including the addition of (coarse and fine) perspective cuts during the solution process. The column labeled *Time factor* $M1/M2 - M1/M3$ refers to the average ratio between the processing time of model $(M1)$ and the processing time of models $(M2)$ and $(M3)$, respectively. The columns labeled *#FPC* and *#CPC* report the average number of fine and coarse perspective cuts, respectively, added during the optimization. The column labeled *Opt. Gap. (%)* report the average optimality gap in percentage. The column labeled *B&B nodes* reports the average number of nodes explored in the branching tree. Finally, the column labeled *Fail* shows the number of times a given formulation fails to solve the instance to optimality within the solution time limit. A file showing the detailed results for the $(M1)$ and $(M3)$ formulations is available at https://github.com/MarioBasallo/IHND.git.

Formulation $(M2)$ provides an approximate solution faster than other formulations in most of the instances. On average, this formulation is 7.5 (6.8) times faster than formulation $(M1)$, for the AP (COL) dataset. For its part, formulation $(M3)$ is 1.3 (1.0) times faster than formulation $(M1)$ for the AP (COL) dataset. It is noted that this factor tends to increase as the size of the network increases. Formulation $(M3)$ can solve a higher number of instances to optimality than formulation $(M1)$ for the AP dataset, but has a similar performance to formulation $(M1)$ for the COL dataset. This suggests that the consideration of coarse perspective cuts in formulation $(M3)$ has a positive impact on reducing the computational time and increasing the number of instances solved to optimality compared to formulation $(M1)$.

The total time used for the lazy constraints function (*Cut time*) is notably smaller for Formulation $(M2)$, which only considers the addition of coarse perspective cuts. This is due to two reasons. On the one hand, for formulation $(M2)$ there is a known closed-form expression of the $\alpha$-level set and its first derivative, which can be computed efficiently. On the other hand, coarse perspective cuts do not need to be defined for each commodity, contrary to fine perspective cuts. In general, formulation $(M1)$ requires on average 2.8(3.2) times more cuts than Formulation $(M3)$, for the AP(COL) dataset.

Formulation $(M3)$ is an extended formulation that introduces coarse perspective cuts to approximate the feasible set of probabilistic service level constraints. Then fine perspective cuts are introduced to improve this approximation. The total number of binary variables in formulation $(M3)$ is twice the number of binary variables in formulation $(M1)$. Despite this difference in the number of binary variables, the results show that formulation $(M3)$ outperforms formulation $(M1)$ for the AP dataset, and it has a similar performance for the COL dataset. This could be explained in different ways. The homothetic approximation of probabilistic service level constraints provides a high approximation accuracy. Also, the cut-generation process is more efficient since adding fine perspective cuts is avoided at the beginning of the optimization.

### 7.4. Managerial insights

Tables 6 and 7 present the characteristics of the intermodal hub network for different probabilistic service level requirements for the AP and COL datasets. Note that for higher service levels (e.g. 95% and 99%), only a few or no hubs are located, and the total cost is primarily comprised of direct transportation costs, as will be discussed later. The high probabilistic service level requirements in such scenarios allow only a small fraction of the total flow to utilize intermodal transportation options, thereby limiting potential savings achievable through combined rail-road transportation.

Table 4: Summary of performance of cutting plane algorithms for the AP dataset.

| Nodes | | Solution time | | | | Cut | Time factor | | | Opt. | B&B | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (# instances) | Model | Min | Avg | Max | St. Dev. | Time | $M1/M2 - M1/M3$ | # FPC | # CPC | Gap(%) | nodes | Fail |
| | (M1) | <1 | **<1** | 1.3 | < 1 | <1 | | 24 | 0 | 0.00 | 19 | 0 |
| 10 (60) | (M2) | <1 | **<1** | 1.8 | < 1 | <1 | $1.1 - 1.3$ | 0 | 24 | 0.00 | 23 | 0 |
| | (M3) | <1 | **<1** | 1.6 | < 1 | <1 | | 20 | 5 | 0.00 | 26 | 0 |
| | (M1) | <1 | **178** | 648 | 192 | 2.9 | | 226 | 0 | 0.00 | 846 | 0 |
| 20 (60) | (M2) | <1 | **126** | 360 | 108 | <1 | $1.4 - 1.4$ | 0 | 68 | 0.00 | 176 | 0 |
| | (M3) | <1 | **142** | 478 | 129 | 1.9 | | 93 | 21 | 0.00 | 464 | 0 |
| | (M1) | <1 | **10,503** | 79,938 | 17,790 | 15.3 | | 765 | 0 | 0.02 | 11,337 | 2 |
| 25 (60) | (M2) | 1.5 | **2,102** | 11,004 | 2,764 | <1 | $6.8 - 5.0$ | 0 | 109 | 0.00 | 557 | 0 |
| | (M3) | 1.7 | **3,000** | 17,751 | 4,546 | 7.7 | | 163 | 39 | 0.00 | 2,226 | 0 |
| | (M1) | 2.7 | **27,025** | 83,405 | 33,863 | 160.0 | | 1,367 | 0 | 4.26 | 3,622 | 39 |
| 40 (60) | (M2) | 1.1 | **16,000** | 69,732 | 17,515 | 1.3 | $20.4 - 19.3$ | 0 | 246 | 11.33 | 470 | 17 |
| | (M3) | 4.3 | **24,452** | 73,959 | 24,196 | 68.0 | | 249 | 68 | 3.82 | 2,616 | 21 |

Table 5: Summary of performance of cutting plane algorithms for the COL dataset.

| Nodes | Model | Solution time | | | | Cut | Time factor | # FPC | # CPC | Opt. | B&B | Fail |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (# instances) | | Min | Avg | Max | St. Dev. | Time | $M1/M2 - M2/M1$ | | | Gap(%) | nodes | |
| 10 (60) | (M1) | <1 | **2.3** | 9.0 | 1.9 | <1 | 0.9 – 0.7 | 121 | 0 | 0.00 | 363 | 0 |
| | (M2) | <1 | **3.3** | 9.7 | 2.7 | <1 | | 0 | 89 | 0.00 | 223 | 0 |
| | (M3) | <1 | **5.1** | 18.0 | 4.6 | <1 | | 68 | 26 | 0.00 | 843 | 0 |
| 20 (60) | (M1) | 5 | **677** | 2,803 | 653 | 3.1 | 1.3 – 0.9 | 528 | 0 | 0.00 | 3,368 | 0 |
| | (M2) | 2 | **533** | 1,413 | 420 | <1 | | 0 | 155 | 0.00 | 820 | 0 |
| | (M3) | 2 | **2,079** | 35,031 | 5,723 | 2.2 | | 140 | 56 | 0.28 | 22,211 | 1 |
| 25 (60) | (M1) | 26 | **6,790** | 34,001 | 7,580 | 11.5 | 1.5 – 1.2 | 923 | 0 | 0.00 | 6,108 | 0 |
| | (M2) | 22 | **4,785** | 23,609 | 5,189 | <1 | | 0 | 174 | 0.00 | 1,239 | 0 |
| | (M3) | 19 | **6,630** | 24,057 | 6,648 | 6.8 | | 185 | 73 | 0.00 | 5,235 | 0 |
| 40 (60) | (M1) | 8,014 | **27,648** | 66,714 | 16,008 | 123.3 | 1.7 – 1.3 | 1,686 | 0 | 2.05 | 2,190 | 37 |
| | (M2) | 4 | **25,605** | 80,178 | 25,042 | <1 | | 0 | 107 | 2.97 | 244 | 34 |
| | (M3) | 1,808 | **27,371** | 84,351 | 24,966 | 82.6 | | 198 | 123 | 2.77 | 1,920 | 37 |

**Table 6** **Impact of probabilistic service level constraints on the hub network characteristics - AP dataset.**

| $r$ | $100\alpha$ | # hubs | Network service level (%) | Modal shift (%) | Network connectivity (%)* | Total Cost |
|---|---|---|---|---|---|---|
| 2 (Tight) | 80 | 3.7 | 87.0 | 27.3 | 93.6 | 321,604 |
| | 85 | 3.6 | 89.7 | 21.8 | 86.6 | 340,897 |
| | 90 | 2.8 | 92.5 | 12.4 | 91.6 | 372,645 |
| | 95 | 0 | 0 | 0 | 0 | 394,492 |
| | 99 | 0 | 0 | 0 | 0 | 394,492 |
| 3 (Moderate) | 80 | 5.1 | 93.6 | 58.4 | 100.0 | 251,724 |
| | 85 | 5.0 | 94.9 | 53.5 | 99.4 | 261,407 |
| | 90 | 5.2 | 96.1 | 48.7 | 100.0 | 272,152 |
| | 95 | 4.1 | 97.7 | 37.5 | 99.2 | 295,072 |
| | 99 | 2.8 | 99.4 | 14.7 | 85.9 | 359,508 |
| 4 (Loose) | 80 | 4.8 | 96.3 | 70.7 | 100.0 | 233,632 |
| | 85 | 4.7 | 96.9 | 66.8 | 99.7 | 236,585 |
| | 90 | 5.1 | 98.0 | 64.3 | 98.5 | 242,604 |
| | 95 | 5.1 | 98.9 | 58.6 | 100.0 | 252,331 |
| | 99 | 4.8 | 99.7 | 42.8 | 97.6 | 283,730 |

* Network connectivity refers to the fraction of inter-hub links that are activated.

**Table 7** **Impact of probabilistic service level constraints on the hub network characteristics - COL dataset.**

| $r$ | $100\alpha$ | # hubs | Network service level (%) | Modal shift (%) | Network connectivity (%) | Total cost |
|---|---|---|---|---|---|---|
| 2 (Tight) | 80 | 2.8 | 85.7 | 15.7 | 100.0 | 530,533 |
| | 85 | 2.3 | 89.3 | 12.4 | 100.0 | 539,103 |
| | 90 | 2.1 | 92.3 | 9.3 | 100.0 | 548,805 |
| | 95 | 0.0 | 0.0 | 0.0 | 0.0 | 550,847 |
| | 99 | 0.0 | 0.0 | 0.0 | 0.0 | 550,847 |
| 3 (Moderate) | 80 | 6.2 | 91.5 | 41.2 | 92.1 | 485,317 |
| | 85 | 6.0 | 93.5 | 37.8 | 89.0 | 491,525 |
| | 90 | 5.3 | 95.4 | 30.8 | 89.5 | 500,772 |
| | 95 | 3.7 | 97.2 | 21.1 | 97.8 | 516,561 |
| | 99 | 2.3 | 99.4 | 11.0 | 100.0 | 545,225 |
| 4 (Loose) | 80 | 6.3 | 93.9 | 47.4 | 92.5 | 470,949 |
| | 85 | 6.2 | 95.6 | 45.8 | 91.8 | 474,253 |
| | 90 | 6.3 | 97.0 | 44.5 | 89.7 | 478,253 |
| | 95 | 6.1 | 98.5 | 40.1 | 90.0 | 486,497 |
| | 99 | 4.6 | 99.6 | 25.8 | 90.8 | 509,185 |

**7.4.1. Effect of varying service levels on total cost** We analyze the effect of varying the service level requirement on the total cost. In Figure 6, dashed lines depict the total cost for the solution when probabilistic service level constraints are dropped and only simple capacity constraints (31d) are considered. The capacity parameter $\Lambda_m$ is computed as in Equation (20) and is affected by the choice of the service level $\alpha$, as discussed before. Solid lines correspond to the case when the model is solved including the probabilistic service level constraints. Figure 6 shows that the total cost increases monotonically as the service level increases. Similarly, the total cost for the model with simple capacity constraints increases monotonically with $\alpha$, which is a consequence of the fact that parameter $\Lambda_m$ is affected by the choice of the service level. It is noted that the total cost for the model without probabilistic service level constraints is smaller and it grows at a slower rate than the total cost of the model with probabilistic service level constraints. The reason for this is that probabilistic service level constraints are more restrictive than simple capacity constraints.

For $r = 3$ and $r = 4$, the total cost increases exponentially as the service level $\alpha$ increases. This behavior is more pronounced for $r = 3$ because, in this scenario, probabilistic service level constraints are tighter, diminishing the likelihood of achieving intermodal transport savings through combined rail-road transportation.

Conversely, when $r = 2$, the total cost exhibits a $S$-shaped curve. Initially, there is exponential growth, but at a certain point, the cost starts to increase at a decreasing rate, eventually stabilizing at a limiting value. This limiting value represents the cost when the demand is entirely satisfied using direct transportation. In this case, probabilistic service level constraints are too restrictive, making it unviable to use a more economical combined transport alternative (see also Figure 8).

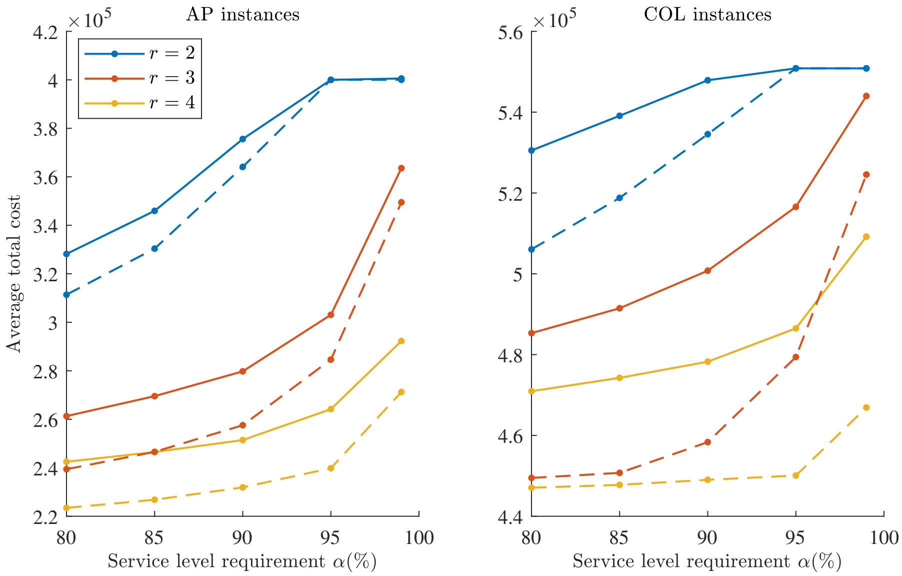


**Figure 6** **Average total cost. Solid lines are the results including probabilistic service level constraints. Dashed lines are the results without probabilistic service level constraints.**

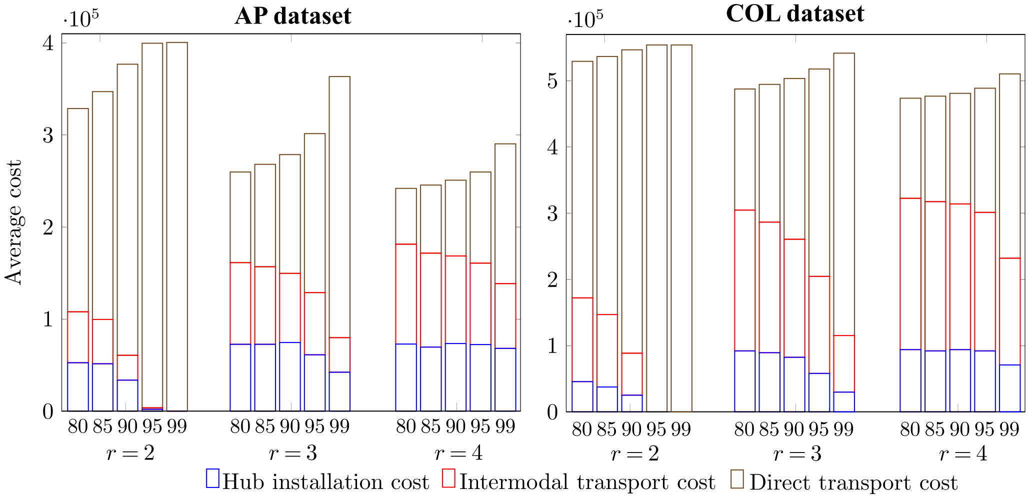


**Figure 7** **Components of the total cost.**

**7.4.2. Effect of varying service levels on cost components and modal shift:** Figure 7 depicts the effect of varying service levels on cost components, i.e., hub installation cost, intermodal transportation cost, and direct transportation cost. As shown in the figures, as the service level increases, the direct transportation costs increase, the intermodal transportation costs decrease, and the hub installation cost decreases. However, the extent of changes in each cost component is comparatively less pronounced when $r = 4$. In this scenario, the average investment in hub facilities remains relatively stable across different service levels, except for $\alpha = 0.99$. An explanation for this behavior is that the average number of open hubs is relatively stable over all instances, as shown in Tables 6 and 7. For $r = 4$, the intermodal transportation costs surpass the direct transportation costs, except when $\alpha = 0.99$. This implies that when the service time requirement is high, i.e., high values of $r$, the cost structure is expected to be less sensitive to changes in the service level and the intermodal transportation costs constitute a significant fraction of the total costs.

The *modal shift* represents the fraction of flow captured by the intermodal transportation (rail-road transportation) and calculated as $\left(\sum_{k\in K}\sum_{a\in A_k} w_k x_{ka}\right) \times 100\%/\sum_{k\in K} w_k$. Figure 8 shows the effect of increasing service level requirements on the modal shift. As expected, as the service level requirement increases, the volume of flows through intermodal transportation decreases. For $r = 3$ and $r = 4$, the modal shift decreases at a diminishing rate when the service level $\alpha$ increases. This decrease in modal shift is more pronounced at elevated service levels and strict service time requirements. Conversely, for $r = 2$, the modal shift curve exhibits an inverted $S$-shape pattern. In this scenario, the modal shift rapidly decreases up to a certain point, after which it decreases more gradually, reaching a modal shift of zero where the intermodal transportation alternative becomes non-viable. The modal shift demonstrates an opposite behavior compared to the total cost.

Figure 9 shows the minimum and average network service levels for formulations with and without probabilistic service level constraints. When probabilistic service level constraints are considered,

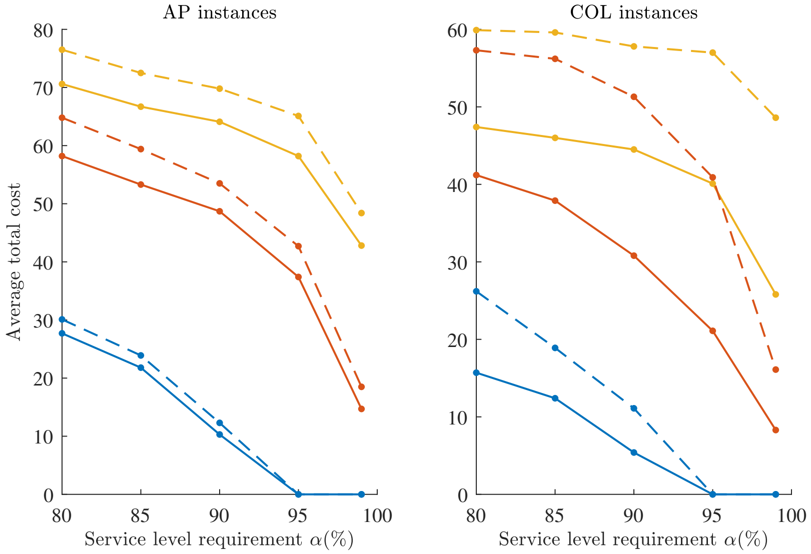


Figure 8 **Average modal shift. Solid lines are the results with probabilistic service level constraints. Dashed lines are the results without probabilistic service level constraints.**

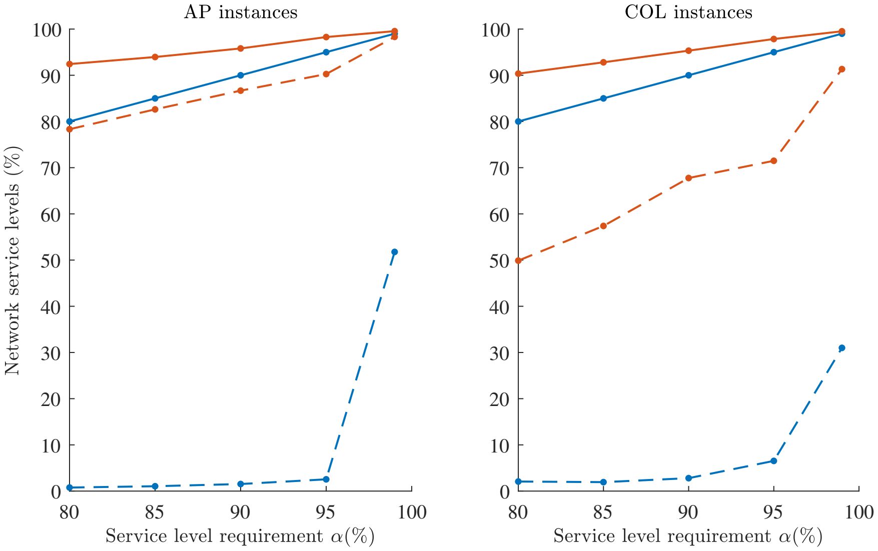


Figure 9 **Minimum and average network service levels. Solid lines are the results with probabilistic service level constraints. Dashed lines are the results without probabilistic service level constraints.**

the minimum service level achieved by the network either coincides with or is higher than the required service level $\alpha$, with a tolerance error of $\varepsilon = 10^{-6}$. The actual service level attained by most inter-hub links in the network exceeds $\alpha$. Some inter-hub links may offer a high service level approaching 100%, particularly those connecting highly capacitated hubs. Note that a hub's total capacity is limited by the hub with the smallest capacity to which the first hub is connected. In contrast, when probabilistic service level constraints are relaxed, and the model is solved using simple capacity constraints, the network's quality of service deteriorates significantly. In such instances, the network service level can fall below 10% for certain inter-hub links. The service quality of networks designed without probabilistic service level constraints experiences a considerable decline, particularly for the COL dataset.

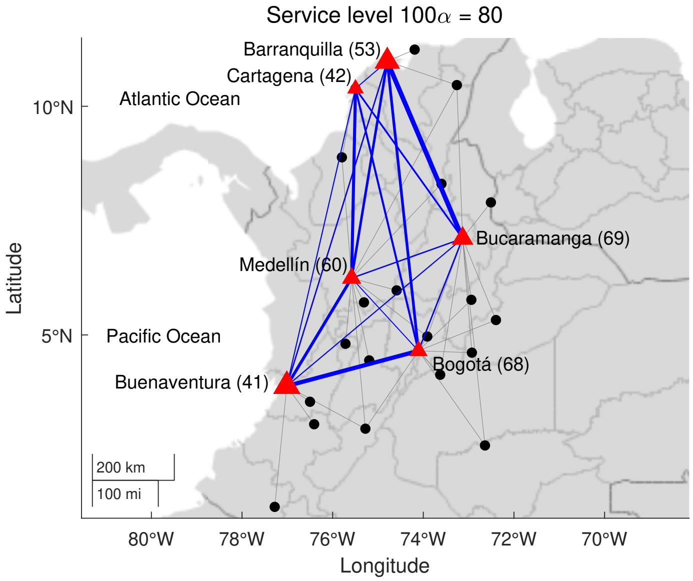


**Figure 10** **Intermodal hub network for the 25-3 instance of the COL dataset with** $100\alpha = 80\%$.

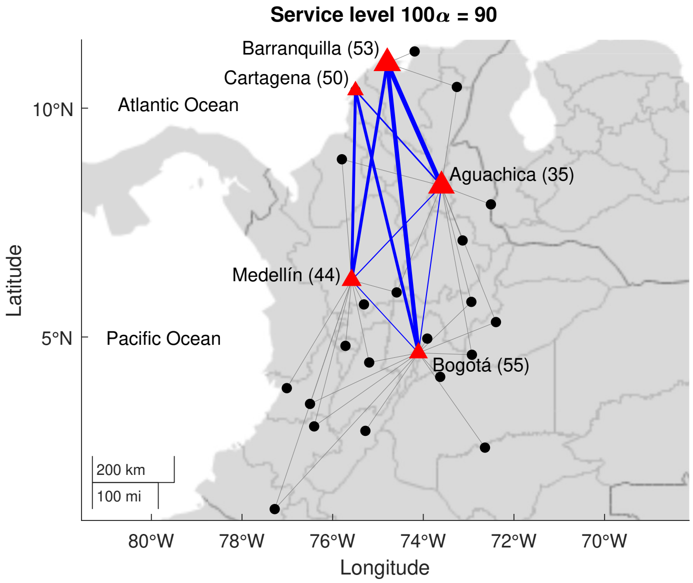


**Figure 11** **Intermodal hub network for the 25-3 instance of the COL dataset with** $100\alpha = 90\%$.

**7.4.3. Effect of varying service levels on the configuration of intermodel hub network:** Figures 10 to 12 depict the configuration of the intermodal hub networks for service level requirements of 80%, 90%, and 99%, respectively with the 25-3 COL instance (where $r = 3$). Note

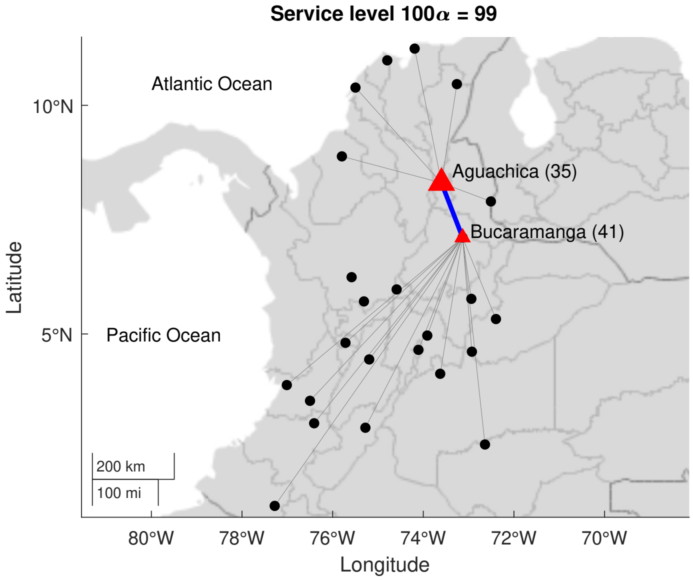


**Figure 12** **Intermodal hub network for the 25-3 instance of the COL dataset with** $100\alpha = 99\%$.

that at a service level of 80%, the inter-hub network is fully connected, with key inter-hub links including Barranquilla-Bucaramanga and Buenaventura-Bogotá. Conversely, at a service level of 90%, the optimal inter-hub network is incomplete, excluding the Barranquilla-Cartagena link, even though it is feasible. In this case, the important inter-hub links shift to Barranquilla-Aguachica and Barranquilla-Bogotá.

In scenarios with lower service levels, hubs are dispersed over a broad geographic area. As the service level increases, hub locations tend to concentrate within a smaller geographic region, specifically the north-central part of Colombia. Several optimal inter-hub networks appear incomplete across different instances, with the number of missing arcs typically being small relative to the total number of feasible hub arcs.

At a service level of 99%, only two hubs (Aguachica and Bucaramanga) are open, resulting in a substantial reduction in modal shift. In this scenario, some origin nodes, located in the central region of the country, do not use any hub at all to satisfy their demand. The numbers in parenthesis in front of the node selected as a hub correspond to the hub utilization, computed as $100\left(\sum_{k\in K}\sum_{a\in\mathcal{A}_k, m\in a} w_k x_{ka}\right)/\mu_m$, for hub $m$. In general, hub capacity utilization is less than 70%, and it tends to decrease as the service level increases.

## 8. Conclusion

We studied the intermodal hub network design problem with probabilistic service level constraints on the total service time of the intermodel transportation route. The probabilistic service level constraints ensure that commodities using the intermodal routes are delivered within a prescribed total service time (transportation time on the three arcs plus the waiting time and processing time at hubs), with a probability greater or equal to a threshold value. To capture the variability in the arrival and service processes of commodities and the resulting delays due to congestion at hubs, we modeled hubs as spatially distributed M/M/1 queues. Several important properties of the total sojourn time and the total service time distribution were analyzed to simplify the probabilistic service level constraints. The resulting formulation contains many logical constraints with indicator variables, which are formulated using perspective cuts. We also propose a homothetic outer approximation for logical constraints and valid inequalities providing a notorious computational advantage over the traditional perspective cuts. The proposed formulations were solved using a cutting plane algorithm. Through sensitivity analyses, we analyze the effect of varying service level requirements on the optimal network configurations. We demonstrate that optimal network configuration that accounts for the service levels may differ significantly from the one that does not consider the service level constraints.

## References

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

Alumur SA, Campbell JF, Contreras I, Kara BY, Marianov V, O'Kelly ME (2021) Perspectives on modeling hub location problems. *European Journal of Operational Research* 291(1):1 – 17.

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(1-2).

Basallo-Triana M, Bravo-Bastidas J, Contreras I, Cordeau JF, Vidal-Holguín C (2023) Intermodal hub network design with generalized capacity constraints and non-synchronized train–truck operations. *Transportation Research Part B: Methodological* 174:102770.

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.

Boysen N, Fliedner M, Jaehn F, Pesch E (2013) A survey on container processing in railway yards. *Transportation Science* 47(3):312–329.

Contreras I, Cordeau JF, Laporte G (2011) Stochastic uncapacitated hub location. *European Journal of Operational Research* 212(3):518–528.

Contreras I, O'Kelly M (2019) *Hub Location Problems*, 327–363 (Cham: Springer International Publishing).

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

Ernst A, Krishnamoorthy M (1996) Efficient algorithms for the uncapacitated single allocation-hub median problem. *Location Science* 4(3):139–154.

Frangioni A, Gentile C (2006) Perspective cuts for a class of convex 0-1 mixed integer programs. *Mathematical Programming* 106(2):225 – 236.

Gao Y, Qin Z (2016) A chance constrained programming approach for uncertain p-hub center location problem. *Computers and Industrial Engineering* 102:10–20.

Ghaffarinasab N (2018) An efficient matheuristic for the robust multiple allocation p-hub median problem under polyhedral demand uncertainty. *Computers and Operations Research* 97:31–47.

Ghaffarinasab N (2022) Stochastic hub location problems with Bernoulli demands. *Computers and Operations Research* 145.

Ghaffarinasab N, CavuS O, Kara BY (2023) A mean-CVaR approach to the risk-averse single allocation hub location problem with flow-dependent economies of scale. *Transportation Research Part B: Methodological* 167:32 – 53.

Hijazi H, Bonami P, Cornuéjols G, Ouorou A (2012) Mixed-integer nonlinear programs featuring ”on/off” constraints. *Computational Optimization and Applications* 52(2):537 – 558.

Hu QM, Hu S, Wang J, Li X (2021) Stochastic single allocation hub location problems with balanced utilization of hub capacities. *Transportation Research Part B: Methodological* 153:204–227.

Ibe OC (2009) *Introduction to Probability Models* (San Diego, CA, USA: Elsevier Academic Press), first edition.

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

Jayaswal S, Vidyarthi N (2023) Multiple allocation hub location with service level constraints for two shipment classes. *European Journal of Operational Research* 309(2):634–655.

Marianov V, Serra D (2003) Location models for airline hubs behaving as m/d/c queues. *Computers and Operations Research* 30(7):983–1003.

Martins de Sá E, Morabito R, de Camargo R (2018a) Benders decomposition applied to a robust multiple allocation incomplete hub location problem. *Computers and Operations Research* 89:31–50.

Martins de Sá E, Morabito R, de Camargo R (2018b) Efficient Benders decomposition algorithms for the robust multiple allocation incomplete hub location problem with service time requirements. *Expert Systems with Applications* 93:50–61.

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

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

Mohammadi M, Jolai F, Tavakkoli-Moghaddam R (2013) Solving a new stochastic multi-mode p-hub covering location problem considering risk by a novel multi-objective algorithm. *Applied Mathematical Modelling* 37(24):10053–10073.

Mohammadi M, Jula P, Tavakkoli-Moghaddam R (2017) Design of a reliable multi-modal multi-commodity model for hazardous materials transportation under uncertainty. *European Journal of Operational Research* 257(3):792–809.

Ross SM (2010) *Introduction to Probability Models* (San Diego, CA, USA: Elsevier Academic Press), tenth edition.

Shahabi M, Unnikrishnan A (2014) Robust hub network design problem. *Transportation Research Part E: Logistics and Transportation Review* 70:356–373.

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):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(6):1446–1470.

Taherkhani G, Alumur S, Hosseini M (2021) Robust stochastic models for profit-maximizing hub location problems. *Transportation Science* 55(6):1322–1350.

Zetina C, Contreras I, Cordeau JF, Nikbakhsh E (2017) Robust uncapacitated hub location. *Transportation Research Part B: Methodological* 106:393–410.

## Appendix A: Proof of Proposition 1

To simplify our analysis, we introduce a change of variables and rewrite the total sojourn time distribution function accordingly. Let $x = (\mu_{a_1} - \lambda_{a_1})v_a$ and $y = (\mu_{a_2} - \lambda_{a_2})v_a$, where $x > 0$ and $y > 0$, then, Equation (7) is transformed to:

$$\tilde{W}(x,y) = \begin{cases} 1 - \frac{xe^{-y/v_a} - ye^{-x/v_a}}{x/v_a - y/v_a}, & \text{for } x \neq y, \\ 1 - e^{-x} - xe^{-x}, & \text{for } x = y, \end{cases}$$
$$= \begin{cases} 1 - \frac{xe^{-y} - ye^{-x}}{x-y}, & \text{for } x \neq y, \\ 1 - e^{-x} - xe^{-x}, & \text{for } x = y. \end{cases}$$

### Proof of Property 1

It is easily verified that $\tilde{W}(x,y) = \tilde{W}(y,x)$, then $\tilde{W}(x,y)$ is symmetric along the line $x = y$. We can conclude that the total sojourn time distribution $W_{V_a}(v_a|\lambda_{a_1}, \lambda_{a_2})$ is symmetric along the line $\mu_{a_1} - \lambda_{a_1} = \mu_{a_2} - \lambda_{a_2}$.

### Proof of Property 2

For a given strictly positive constant $c$, we have

$$\lim_{(x,y)\to(c,c)} \tilde{W}(x,y) = 1 - e^{-c} - ce^{-c} = \tilde{W}(c,c),$$

where the previous limit is computed assuming that $x \neq y$. This implies that $\tilde{W}(x,y)$ is continuous and so is the total sojourn time distribution $W_{V_a}(v_a|\lambda_{a_1}, \lambda_{a_2})$.

The partial derivatives of $\tilde{W}$ are

$$\frac{\partial \tilde{W}}{\partial x} = \frac{y}{(x-y)^2}\left[e^{-y} - e^{-x}(1+x-y)\right], \quad \text{and} \quad \frac{\partial \tilde{W}}{\partial y} = \frac{x}{(x-y)^2}\left[e^{-x} + e^{-y}(-1+x-y)\right], \tag{39}$$

which have a finite limit of $ce^{-c}/2$ when $(x,y) \to (c,c)$. In particular, we are interested in the derivative of the level sets of $\tilde{W}$, which is computed as:

$$-\frac{\partial \tilde{W}/\partial x}{\partial \tilde{W}/\partial y} = \frac{y}{x}\frac{e^{-y} - e^{-x}(1+x-y)}{e^{-x} + e^{-y}(-1+x-y)}.$$

This derivative has a limiting value of $-1$ at singularity points, as it is expected by the symmetry of $\tilde{W}$.

### Proof of property 3

First, we prove that $\tilde{W}(x,y)$ is a strictly increasing function. Consider the case $x = y$, we have $\partial \tilde{W}/\partial x = xe^{-x} > 0$, so $\tilde{W}$ is strictly increasing on the line $x = y$.

Now consider that $x \neq y$, then $\partial \tilde{W}/\partial x$, given in Equation (39), is non-negative if and only if $e^{-y} + e^{-x}(y - x - 1) \geq 0$, which is equivalent to $e^{x-y} \geq x - y + 1$. Given that $e^{x-y}$ is convex and $x - y + 1$ is tangent to $e^{x-y}$, then the inequality holds, with a strict equality when $x = y$. Moreover, it is known that $\tilde{W}$ is strictly increasing at $x = y$, hence $\partial \tilde{W}/\partial x > 0$. An analogous procedure is followed to show that $\partial \tilde{W}/\partial y > 0$. According to this $\tilde{W}$ is strictly increasing. Finally, we have that $\partial W_{V_a}/\partial \lambda_{a1} = (\partial \tilde{W}/\partial x)(\partial x/\partial \lambda_{a_1}) + (\partial \tilde{W}/\partial y)(\partial y/\partial \lambda_{a_1}) = -v_a \partial \tilde{W}/\partial x < 0$, and analogously it is shown that $\partial W_{V_a}/\partial \lambda_{a2} < 0$. We conclude that the total sojourn time distribution $W_{V_a}(v_a|\lambda_{a_1}, \lambda_{a_2})$ is strictly decreasing.

## Appendix B: Concavity analysis of the total sojourn time distribution

We follow the change of variables suggested in Appendix A. Given that $x$ and $y$ are an affine transformation of the original variables $\lambda_{a_1}$ and $\lambda_{a_2}$, the concavity of the total sojourn time distribution is not altered after the transformation. As a consequence of the transformation, $\tilde{W}(x,y)$ is a strictly increasing function. Now, the proof of concavity for $\tilde{W}(x,y)$ is equivalent to the proof of convexity for the function:

$$f(x,y)=\begin{cases}\frac{xe^{-y}-ye^{-x}}{x-y}, & \text{for } x>0, y>0, x\neq y,\\ e^{-x}+xe^{-x}, & \text{for } x>0, y>0, x=y.\end{cases}$$

The convexity of $f(x,y)$ can be assessed by determining whether the corresponding Hessian matrix $H_f(x,y)$ is positive (semi)definite in the domain of $f$. To this end, we use the leading principal minors criteria. Then, $H_f$ is positive (semi)definite if all its leading principal minors are non-negative. The minor of order 1 of $H_f$ is:

$$|H_f(x,y)|_1=\frac{ye^{-x-y}\left\{2e^x-\left[(1+x-y)^2+1\right]e^y\right\}}{(x-y)^3}.$$

Since $f$ is symmetric, it is enough to show that $|H_f(x,y)|_1$ is non-negative in the region defined by $x\geq y, x>0$, and $y>0$. According to this, $|H_f(x,y)|_1$ is non-negative if and only if $2e^x-\left[(1+x-y)^2+1\right]e^y\geq 0$, which is equivalent to $2e^{x-y}-1-(1+x-y)^2\geq 0$. Let $z=x-y\geq 0$, the left-hand side of the previous inequality becomes $2e^z-1-(1+z)^2$. This expression has a minimum value of 0, which is obtained when $z=0$ or $x=y$. This result can be verified using the first and second derivative criteria. Then, the inequality holds, and we conclude that $|H_f(x,y)|_1$ is non-negative everywhere in the domain of $f$.

On the other hand, the minor of order 2 of $H_f$ is:

$$|H_f(x,y)|_2=\frac{e^{-2(x+y)}\left\{\begin{matrix}-e^{2y}\left[(1+x)^2-2xy\right]-e^{2x}\left[(1+y)^2-2xy\right]+\\ e^{x+y}\left[2+2x+2y-2xy-(x-y)^2xy\right]\end{matrix}\right\}}{(x-y)^4}.$$

In this case, $|H_f(x,y)|_2$ can be negative in the domain of $f$, as it is shown in Figure 13. According to this, $f$ is not convex everywhere in its domain. It is of interest to explicitly define a region where $f$ is convex. Note that the second leading principal minor is non-negative if and only if:

$$-e^{2y}\left[(1+x)^2-2xy\right]-e^{2x}\left[(1+y)^2-2xy\right]+e^{x+y}\left[2+2x+2y-2xy-(x-y)^2xy\right]\geq 0. \tag{40}$$

We conjecture that the previous inequality is satisfied in the region $y\geq 1/x$ and $x>0$. We show that $|H_f(x,y)|_2$ is non-negative along the curve $y=1/x$. Replacing $y=1/x$ in Inequality (40) and simplifying we get:

$$\frac{e^{-\frac{1}{x}}\left(e^x-e^{1/x}x^2\right)\left[e^{x-\frac{1}{x}}\left(x-\frac{1}{x}-2\right)+x-\frac{1}{x}+2\right]}{x^3}\geq 0. \tag{41}$$

By the symmetry of $f$ and given that $y\geq 1/x$, we note that $x\geq 1$ in the region of interest. In this sense, each factor in the numerator of the left-hand side of Inequality (41) is non-negative, as can be easily verified using the first and second derivative criteria. Then, the inequality holds, and the strict equality is obtained only when $x=1$. This allows us to conclude that $H_f$ is positive (semi)definite on the curve $y=1/x$.

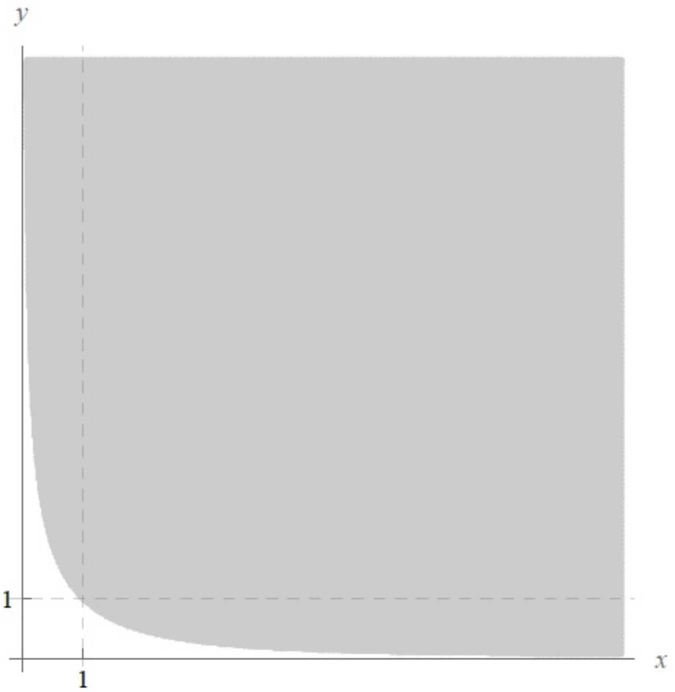


**Figure 13** **The shaded area shows the region where $|H_f(x,y)|_2$ is non-negative, or, equivalently, the region where $H_f$ is positive (semi)definite.**

Based on the conjecture that $f$ is convex (equivalently: $\tilde{W}$ is concave) for $y \geq 1/x$ and $x > 0$, and given that $\tilde{W}(x,y)$ is a strictly increasing function, it is of interest to find the maximum of $\tilde{W}(x,y)$ constrained to $y = 1/x$. We have the following univariate optimization problem:

$$\max_{x>0} \qquad \tilde{W}\left(x, \frac{1}{x}\right) = 1 - \frac{e^{-x} - e^{-\frac{1}{x}}x^2}{1-x^2},$$

which has solution $\tilde{W}^* = 1 - 2/e$, $x^* = 1$. This suggests that a lower bound for the minimum percentile of the total sojourn time distribution at which such distribution remains concave is $(1-2/e)100 \approx 26.4$. Given that, in practical applications, the desired service level $\alpha$ is much greater than $1-2/e$, probabilistic service level constraints can be considered to be convex for relevant practical applications.

## Appendix C: A closed form expression for $f^a$

Consider the case when $\mu_{a_1} - \lambda_{a_1} \neq \mu_{a_2} - \lambda_{a_2}$, from Equation (10) we know that at the $\alpha$-level set the following relation must hold:

$$1 - \frac{(\mu_{a_1} - \lambda_{a_1})e^{-(\mu_{a_2}-\lambda_{a_2})\tau}}{(\mu_{a_1} - \lambda_{a_1}) - (\mu_{a_2} - \lambda_{a_2})} - \frac{(\mu_{a_2} - \lambda_{a_2})e^{-(\mu_{a_1}-\lambda_{a_1})\tau}}{(\mu_{a_2} - \lambda_{a_2}) - (\mu_{a_1} - \lambda_{a_1})} = \alpha.$$

After some algebraic manipulations, the previous expression can be rewritten as:

$$\left[e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\right]\lambda_{a_1} + (\mu_{a_2 - \lambda_{a_2}})e^{-\mu_{a_1}\tau}e^{\lambda_{a_1}\tau} = \left[e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\right]\mu_{a_1} + (1-\alpha)(\mu_{a_2} - \lambda_{a_2}). \quad (42)$$

Let

$$d = e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha),$$
$$b = (\mu_{a_2} - \lambda_{a_2})e^{-\mu_{a_1}\tau},$$
$$c = \left[e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\right]\mu_{a_1} + (1-\alpha)(\mu_{a_2} - \lambda_{a_2}).$$

Equation (42) is rewritten as $d\lambda_{a_1} + be^{\lambda_{a_1}\tau} = c$, and we want to solve for $\lambda_{a_1}$ theis equation. Using algebraic manipulations we have $a\lambda_{a_1} + be^{\lambda_{a_1}t} = c \Longrightarrow \frac{b}{a}te^{\lambda_{a_1}t} = t\frac{c}{a} - \lambda_{a_1}t \Longrightarrow \frac{b}{a}t = \left(\frac{c}{a}t - \lambda_{a_1}t\right)e^{-\lambda_{a_1}t} \Longrightarrow \frac{b}{a}te^{\frac{c}{a}t} = \left(\frac{c}{a} - \lambda_{a_1}\right)te^{\left(\frac{c}{a}-\lambda_{a_1}\right)t}$. Taking the Lambert W function to both sides of the previous equation we obtain $\mathcal{W}\left(\frac{b}{a}te^{\frac{c}{a}t}\right) = \left(\frac{c}{a} - \lambda_{a_1}\right)t$. After solving for $\lambda_{a_1}$ we obtain

$$\lambda_{a_1} = f^a(\lambda_{a_2}) = \frac{c}{d} - \frac{1}{\tau}\mathcal{W}_r\left(\tau\frac{b}{d}e^{\tau\frac{c}{d}}\right), \tag{43}$$

where $\mathcal{W}_r(\cdot)$ is the Lambert W function evaluated at branch $r \in \{-1, 0\}$. Selecting the appropriate branch in the Lambert W function is not trivial. The criterion to select $r$ is the analysis of the singularity that occurs at the $\alpha$-level set of the total sojourn time distribution. Let $(\lambda^*_{a_1}, \lambda^*_{a_2})$ be such a singularity. The following equations must be satisfied at the $\alpha$-level set:

$$1 - e^{-(\mu_{a_1} - \lambda^*_{a_1})\tau} - (\mu_{a_1} - \lambda^*_{a_1})\tau e^{-(\mu_{a_1} - \lambda^*_{a_1})\tau} = \alpha,$$
$$1 - e^{-(\mu_{a_2} - \lambda^*_{a_2})\tau} - (\mu_{a_2} - \lambda^*_{a_2})\tau e^{-(\mu_{a_2} - \lambda^*_{a_2})\tau} = \alpha.$$

The solution for $\lambda^*_{a_1}$ and $\lambda^*_{a_2}$ in previous equations is:

$$\lambda^*_{a_1} = \frac{1}{\tau}\left[1 + \mu_{a_1}\tau + \mathcal{W}_{-1}\left(\frac{\alpha - 1}{e}\right)\right], \tag{44}$$
$$\lambda^*_{a_2} = \frac{1}{\tau}\left[1 + \mu_{a_2}\tau + \mathcal{W}_{-1}\left(\frac{\alpha - 1}{e}\right)\right]. \tag{45}$$

To understand why branch $\mathcal{W}_{-1}$ should be used, note that the first equation can be written as $(\lambda^*_{a_1} - \mu_{a_1})\tau = 1 + \mathcal{W}_{-1}[(\alpha - 1)/e]$, by Equation (8), the left-hand side of such equation is strictly negative, so it is the right-hand side. Then, we must have $\mathcal{W}_r[(\alpha - 1)/e] < -1$. Given that $-1/e < (\alpha - 1)/e < 0$, the previous inequality is satisfied only when the branch $r = -1$ is considered. According to the previous analysis, we have found that the correct branch selection in the Lambert W function in Equation (43) can be done according to the following rule:

$$r = \begin{cases} -1, & \text{if } \lambda_{a_2} \leq \lambda^*_{a_2}, \\ 0, & \text{if } \lambda_{a_2} > \lambda^*_{a_2}. \end{cases}$$

Function $f^a$ has asymptotes at $\lambda_{a_1} = \ln(1-\alpha)/\tau + \mu_{a1}$ and $\lambda_{a_2} = \ln(1-\alpha)/\tau + \mu_{a2}$. An explicit expression for $f^a$ and its first derivative is

$$f^a(\lambda_{a_2}) = \mu_{a_1} + \frac{(1-\alpha)(\mu_{a_2} - \lambda_{a_2})}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)} - \frac{1}{\tau}\mathcal{W}_r\left\{\frac{(\mu_{a_2} - \lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)}\text{Exp}\left[\frac{(1-\alpha)(\mu_{a_2} - \lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)}\right]\right\}, \tag{46}$$

$$\begin{aligned} f^a_1(\lambda_{a_2}) = {} & \frac{(1-\alpha)^2 - (1-\alpha)\left[1 + (\mu_{a_2} - \lambda_{a_2})\tau\right]e^{-(\mu_{a_2}-\lambda_{a_2})\tau}}{\left[e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\right]^2} + \\ & \frac{\mathcal{W}_r\left\{\frac{(\mu_{a_2}-\lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau}-(1-\alpha)}\text{Exp}\left[\frac{(1-\alpha)(\mu_{a_2}-\lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau}-(1-\alpha)}\right]\right\}}{1 + \mathcal{W}_r\left\{\frac{(\mu_{a_2}-\lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau}-(1-\alpha)}\text{Exp}\left[\frac{(1-\alpha)(\mu_{a_2}-\lambda_{a_2})\tau}{e^{-(\mu_{a_2}-\lambda_{a_2})\tau}-(1-\alpha)}\right]\right\}} \times \\ & \frac{\left\{(1-\alpha) - \left[1 + (\mu_{a_2} - \lambda_{a_2})\tau\right]e^{-(\mu_{a_2}-\lambda_{a_2})\tau}\right\}\left\{e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\left[1 - (\mu_{a_2} - \lambda_{a_2})\tau\right]\right\}}{-\tau(\mu_{a_2} - \lambda_{a_2})\left[e^{-(\mu_{a_2}-\lambda_{a_2})\tau} - (1-\alpha)\right]^2}. \end{aligned} \tag{47}$$

## Appendix D: Proof of Proposition 4

Consider parameter $\tau$ and let $g(x,y\,|\,\tau)$ be a function of variables $x$ and $y$ defined as follows

$$g(x,y\,|\,\tau)=\frac{xe^{-\tau y}-ye^{-\tau x}}{x-y}.$$

The total service time distribution $W_{V_a}(\tau\,|\,\lambda_{a_1},\lambda_{a_2})$ is related to function $g$ as $W_{V_a}(\tau\,|\,\lambda_{a_1},\lambda_{a_2})=1-g(\mu_{a_1}-\lambda_{a_1},\mu_{a_2}-\lambda_{a_2})$. This corresponds to translations and reflections of function $g$. The origin of coordinates $(0,0)$ is translated to the point $(\mu_{a_1},\mu_{a_2})$. Given that $W_{V_a}$ is an affine transformation of $g$, the homothetic properties of the level sets of $g$ are preserved after the transformation. We will show that the $\alpha$-levels sets of $g(x,y\,|\,\tau)$ and $g(x,y\,|\,\tau')$, for $0<\tau, 0<\tau'\neq\tau$, are homothetic with respect to the origin. This is equivalent to showing that any ray from the origin cuts each $\alpha$-level set at points with equal slope.

Assume that $\tau=a\tau'$, for $a>0$. According to this, we have the following relation

$$\begin{aligned}
g(x,y\,|\,\tau')&=\frac{xe^{-\tau' y}-ye^{-\tau' x}}{x-y},\\
&=\frac{xe^{-(\tau/a)y}-ye^{-(\tau/a)x}}{x-y},\\
&=\frac{(x/a)e^{-\tau(y/a)}-(y/a)e^{-\tau(x/a)}}{(x/a)-(y/a)},\\
&=g(x/a,y/a\,|\,\tau).
\end{aligned}\tag{48}$$

At the $\alpha$-level set we have $g(x,y\,|\,\tau')=g(x/a,y/a\,|\,\tau)=\alpha$. Consider points $(x_0,y_0)$ and $(x_0/a,y_0/a)$, which are collinear to the origin, or are contained in a ray from the origin. If point $(x_0,y_0)$ is in the $\alpha$-level set of $g(x,y\,|\,\tau')$, and point $(x_0/a,y_0/a)$ is in the $\alpha$-level set of $g(x,y\,|\,\tau)$. In this sense, one level set can be interpreted as the projection from the origin of the other level set, and the origin of coordinates is also referred to as the homothetic center of $g$.

To verify that $g(x,y\,|\,\tau')$ and $g(x,y\,|\,\tau)$ have an equal slope at points $(x_0,y_0)$ and $(x_0/a,y_0/a)$, respectively, we use the closed-form expression of the first derivative of $\alpha$-level set of $g$. First, we use Appendix C to obtain the $\alpha$-level set of $g$:

$$h(x\,|\,\tau)=x+\frac{x}{\alpha e^{\tau x}-1}+\frac{1}{\tau}\mathcal{W}\left[\frac{\tau x}{1-\alpha e^{\tau x}}\operatorname{Exp}\left(\frac{\tau x}{1-\alpha e^{\tau x}}\right)\right],$$

which satisfies $h(x\,|\,\tau')/a=h(x/a\,|\,\tau)$, as it is expected from the previous discussion. Differentiating $h$ with respect to $x$ we obtain:

$$h_1(x\,|\,\tau)=(\alpha e^{\tau x}-\tau x-1)\,\frac{\alpha\tau xe^{\tau x}+(\alpha e^{\tau x}-1)\,\mathcal{W}\left[\frac{\tau x}{1-\alpha e^{\tau x}}\operatorname{Exp}\left(\frac{\tau x}{1-\alpha e^{\tau x}}\right)\right]}{\tau x\,(\alpha e^{\tau x}-1)\left\{1+\mathcal{W}\left[\frac{\tau x}{1-\alpha e^{\tau x}}\operatorname{Exp}\left(\frac{\tau x}{1-\alpha e^{\tau x}}\right)\right]\right\}}.$$

It is noted that $h_1(x\,|\,\tau')=h_1(x/a\,|\,\tau)$, in particular, $h_1(x_0\,|\,\tau')=h_1(x_0/a\,|\,\tau)$, as desired.

Given that $W_{V_a}$ transforms $g$ in a way that the origin of coordinates is translated to the point $(\mu_{a_1},\mu_{a_2})$, we conclude that the $\alpha$-level sets of $W_{V_a}(\tau\,|\,\lambda_{a_1},\lambda_{a_2})$ and $W_{V_a}(\tau'\,|\,\lambda_{a_1},\lambda_{a_2})$ are homothetic concerning the point $(\mu_{a_1},\mu_{a_2})$. We refer to point $(\mu_{a_1},\mu_{a_2})$ as the homothetic center of a family of $\alpha$-level sets of $W_{V_a}$ defined for distinct values of parameter $\tau$.

## Appendix E: Numerical evaluation of $f^{ka}$

We describe a numerical method for evaluating the $\alpha$-level set of the total service time distribution $E_\alpha^{S_{\tau_k}}$. First, we need a numerical method to evaluate the Convolution Integral (12). There are many available numerical integration methods in the literature. In this paper, we have used the Gauss-Legendre Quadrature method.

The idea is to compute $\lambda_{a_1}$ for a given value of $\lambda'_{a_2}$, with the condition that $(\lambda_{a_1}, \lambda'_{a_2}) \in E_\alpha^{S_{\tau_k}}(a)$. In this sense, $\lambda_{a_1}$ is the solution of the equation

$$S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda'_{a_2}) = \alpha. \tag{49}$$

Note that the total service time distribution is evaluated at the service time requirement $\tau_k$, which is the value of interest for establishing probabilistic service level constraints. In other words, we want to find the root of the function $S_{T_{ka}}(\tau_k \mid \bullet, \lambda'_{a_2}) - \alpha$. We use Brent's root finding algorithm to solve for $\lambda_{a_1}$ in Equation (49). By the monotonicity property of $S_{T_{ka}}(\tau_k \mid \bullet, \lambda'_{a_2})$, we know that at most one root can be found in the interval $(-\mu_{a_1}, \mu_{a_1})$. No solution may exist for the previous equation. In that case, we say that $\lambda'_{a_2}$ is not feasible.

We use the following notation to write function $f^{ka}(\lambda_{a_2})$:

$$f^{ka}(\lambda_{a_2}) = \mathrm{Root}\left[S_{T_{ka}}(\tau_k \mid \bullet, \lambda_{a_2}) - \alpha\right]. \tag{50}$$

It will also be necessary to compute the first derivative of $f^{ka}$, which is denoted as $f_1^{ka}$. We use the finite difference method to numerically compute such a derivative. Just as $f^a$, $f^{ka}$ has a singularity at point $(\lambda^*_{ka_1}, \lambda^*_{ka_1})$. There is no closed-form expression for such a singularity and it is computed using numerical integration.

## Appendix F: Heuristic algorithm for determining $\tilde{\tau}_{ka}$

The idea of this algorithm is to define $\tilde{\tau}_{ka}$ such that the singularity point of $\tilde{f}^{ka}$ coincides with the singularity point of $f^{ka}$. If this selection of $\tilde{\tau}_{ka}$ leads to a situation where $\tilde{f}^{ka} \geq f^{ka}$ is not satisfied at the extreme points $\left(\lambda^{\max}_{ka_1}(a_2), 0\right)$ and $\left(0, \lambda^{\max}_{ka_2}(a_1)\right)$, then a new $\tilde{\tau}_{ka}$ is recalculated so that the inequality at extreme points is satisfied. The algorithm seems to produce a correct outer approximation, as it is noticed in our numerical results obtaining a quite accurate outer approximation of probabilistic service level constraints.

## Appendix G: Proof of Proposition 8

We first show that $\tilde{\Delta}^{\mathbf{r}}_l \geq 0$. Let $[l] = e(ka)$ and $[l-1] = e(k'a)$. By Equation (33), $\tilde{\delta}^{\mathbf{r}}_{[l]} = \tilde{\delta}^{\mathbf{r}}_{ka} = L_{a_1} - \lambda^{\mathbf{r}}_{ka_1} + |s^{\mathbf{r}}_a|\left(L_{a_2} - \lambda^{\mathbf{r}}_{ka_2}\right) \geq 0$. According to Proposition 4, the point $\left(\lambda^{\mathbf{r}}_{ka_1}, \lambda^{\mathbf{r}}_{ka_2}\right)$ is the projection from the homothetic center $(\mu_{a_1}, \mu_{a_2})$ of the point $\left(\lambda^{\mathbf{r}}_{k'a}, \lambda^{\mathbf{r}}_{k'a_2}\right)$, which is computed as

$$\left(\lambda^{\mathbf{r}}_{ka_1}, \lambda^{\mathbf{r}}_{ka_2}\right) = \left[\mu_{a_1} + \left(\lambda^{\mathbf{r}}_{k'a_1} - \mu_{a_1}\right)\frac{\tilde{\tau}_{k'a}}{\tilde{\tau}_{ka}}, \mu_{a_2} + \left(\lambda^{\mathbf{r}}_{k'a_2} - \mu_{a_2}\right)\frac{\tilde{\tau}_{k'a}}{\tilde{\tau}_{ka}}\right].$$

By Proposition 6, $\tilde{\tau}_{ka} \leq \tilde{\tau}_{k'a}$, then $\lambda^{\mathbf{r}}_{ka_1} \leq \lambda^{\mathbf{r}}_{k'a_1}$ and $\lambda^{\mathbf{r}}_{ka_2} \leq \lambda^{\mathbf{r}}_{k'a_2}$. This implies that $\tilde{\delta}^{\mathbf{r}}_{e(ka)} \geq \tilde{\delta}^{\mathbf{r}}_{e(k'a)}$ and $\tilde{\Delta}^{\mathbf{r}}_{e(ka)} = \tilde{\Delta}^{\mathbf{r}}_{[l]} \geq 0$.

**Algorithm 2:** Heuristic algorithm for computing $\tilde{\tau}_{ka}$

```
Data: $\alpha, \mu_{a_1}, \mu_{a_2}, f^{ka}$.
Compue the singularity point $(\lambda^*_{ka_1}, \lambda^*_{ka_2})$ of $f^{ka}$ using numerical integration (see Section 4);
$\tilde{\tau}_{ka} \leftarrow \frac{1+\mathcal{W}_{-1}[(\alpha-1)/e]}{\lambda^*_{ka_1}-\mu_{a_1}}$;        // Solution for $\tau$ in Equation (44).
Define the function $\tilde{f}^{ka}$ as in Equation (46) as the $\alpha$-level set of $W_{V_a}(\tilde{\tau}_{ka} \mid \lambda_{a_1}, \lambda_{a_2})$;
if $\lambda^{max}_{ka_1}(a_2) > \tilde{f}^{ka}(0)$ or $0 > \tilde{f}^{ka}\left(\lambda^{max}_{ka_2}(a_1)\right)$ then
    if $\lambda^{max}_{ka_1}(a_2) < \lambda^{max}_{ka_2}(a_1)$ then
        Solve for $\tau$ in the Equation $W_{V_a}\left(\tau \mid \lambda^{\max}_{ka_1}(a_2), 0\right) = \alpha$ using numerical methods;
        $\tilde{\tau}_{ka} \leftarrow \tau$;
    else
        Solve for $\tau$ in the Equation $W_{V_a}\left(\tau \mid 0, \lambda^{\max}_{ka_2}(a_1)\right) = \alpha$ using numerical methods;
        $\tilde{\tau}_{ka} \leftarrow \tau$;
    end
end
```

Assume that $v_{[l]} = 1$, for $l = 1, 2, \ldots, l'$. From the perspective cut in Equation (26) we have

$$\begin{aligned} L_{a_1} - \lambda_{a_1} + |s^{\mathbf{r}}_a|(L_{a_2} - \lambda_{a_2}) &\geq \left[L_{a_1} - \lambda^{\mathbf{r}}_{[k']a_1} + |s^{\mathbf{r}}_a|\left(L_{a_2} - \lambda^{\mathbf{r}}_{[k']a_2}\right)\right] v_{[l']} \\ &= \tilde{\delta}^{\mathbf{r}}_{[l']} v_{[l']}, \\ &= \sum_{l=1}^{|\mathscr{L}_a|-1} \tilde{\delta}^{\mathbf{r}}_{[l]} \left(v_{[l]} - v_{[l+1]}\right) + \tilde{\delta}^{\mathbf{r}}_{|\mathscr{L}_a|} v_{|\mathscr{L}_a|}, \\ &= \tilde{\delta}^{\mathbf{r}}_{[1]} v_{[1]} + \sum_{l=2}^{|\mathscr{L}_a|} \left(\tilde{\delta}^{\mathbf{r}}_{[l]} - \tilde{\delta}^{\mathbf{r}}_{[l-1]}\right) v_{[l]}, \\ &= \sum_{l=1}^{|\mathscr{L}_a|} \tilde{\Delta}^{\mathbf{r}}_{[l]} v_{[l]}, \\ &= \sum_{l \in \mathscr{L}_a} \tilde{\Delta}^{\mathbf{r}}_l v_l, \end{aligned}$$

where the second equality is valid due to the incremental nature of binary variables $v_{[l]}$. Reorganizing terms we get

$$\lambda_{a_1} + |s^{\mathbf{r}}_a|\lambda_{a_2} + \sum_{l \in \mathscr{L}_a} \tilde{\Delta}^{\mathbf{r}}_l v_l \leq L_{a_1} + |s^{\mathbf{r}}_a| L_{a_2}. \tag{51}$$

Multiplying $L_{a_1}$ and $L_{a_2}$ by $z_{a_1}$ and $z_{a_2}$, respectively, in the right-hand side of the previous is valid and allows to obtain Inequality (34), as desired.

## Appendix H: The multiple choice formulation

Reconsider incremental variables $v_l$ and introduce new binary variables $u_l$ defined such that $u_{[l]} = v_{[l-1]} - v_{[l]}$, where $u_{[1]} = v_{[1]}$. Using this definition, coarse perspective cuts are defined in terms of variables $u_l$ as follows:

$$\lambda_{a_1} + |s^{\mathbf{r}}_a|\lambda_{a_2} + \sum_{l \in \mathscr{L}_a} \tilde{\delta}^{\mathbf{r}}_l u_l \leq L_{a_1} z_{a_1} + |s^{\mathbf{r}}_a| L_{a_2} z_{a_2}, \qquad \forall a \in \mathscr{B}, \mathbf{r} \in \mathscr{C}_a. \tag{52}$$

Using this change of variables, the multiple-choice formulation is as follows

$$\begin{aligned}
\text{minimize} \quad & \sum_{m\in\mathcal{N}} f_m z_m - \sum_{k\in K}\sum_{a\in\mathcal{A}_k}(C_k - C_{ka})x_{ka} + \sum_{k\in K} C_k && (53a)\\
\text{subject to} \quad & (31b)-(31d), (31f), (31g), (52), && \\
& x_{ka} \le \sum_{l=1}^{\mathbf{Pos}[e(ka)]} v_{[l]}, \qquad \forall k\in K, a\in\mathcal{A}_k, && (53b)\\
& \sum_{l\in\mathcal{L}_a} v_l \le z_m, \qquad \forall a\in\mathcal{B}, m\in a, && (53c)\\
& v_l \in\{0,1\}, \qquad \forall a\in\mathcal{B}, l\in\mathcal{L}_a, && (53d)
\end{aligned}$$

where the expression $\mathbf{Pos}(l)$ denotes the position occupied by the transport path $l$ in the list of transport paths ordered following the homothetic ordering of commodities in strictly distinguishable transport paths.

## Appendix I: Valid inequalities for formulation $(M1)$

We note that coarse perspective cuts (34) are still valid for $(M1)$ in the absence of incremental constraints (see Proposition 9). However, these cuts are not strong enough in this case. It is possible to consider similar incremental constraints for some subsets of commodities.

PROPOSITION 9. *Assume that $\tilde{f}^{ka}$ is a valid outer approximation of $f^{ka}$, then in the absence of incremental constraints, Coarse Perspective Cuts (34) are still valid inequalities for model (M1).*

*Proof* Assume that $y_l = 1$, for $l\in S_a\subseteq\mathcal{L}_a$, and $y_l = 0$, for $l\notin S_a$. Let $[l'] = \arg\max_{[l]\in S_a}\{\tilde{\delta}^{\mathbf{r}}_{[l]}\}$. From the perspective cut in Equation (26) we have

$$L_{a_1} - \lambda_{a_1} + |s^{\mathbf{r}}_a|(L_{a_2} - \lambda_{a_2}) \ge \tilde{\delta}^{\mathbf{r}}_{[l']a} y_{[l']} \ge \sum_{[l]\in S_a}\left(\tilde{\delta}^{\mathbf{r}}_{[l]} - \tilde{\delta}^{\mathbf{r}}_{[l-1]}\right) y_{[l]} = \sum_{l\in\mathcal{L}_a}\tilde{\Delta}^{\mathbf{r}}_l y_l.$$

Reorganizing terms and introducing binary variables $z_{a_1}$ and $z_{a_2}$ as in Proposition 9, we get Inequality (34). □

It is important to note that incremental constraints of the form $y_{[l]}\le y_{[l-1]}$ given in Proposition 7 are not valid for all transport paths. However, similar constraints might be determined for some subsets of transport paths associated with commodities that can be ordered. A non-exhaustive approach to this end is given in the following proposition.

PROPOSITION 10. *If $\tau_k\le\tau_l$ and $G_{T_{ka}}\left(\frac{\tau_k}{\tau_l}x\right)\le G_{T_{la}}(x)$, for $x\in[0,\tau_l]$, then $y_{e(ka)}\le y_{e(la)}$ is a valid inequality for $(M1)$.*

*Proof* The total service time distribution for commodities $k$ and $l$ is:

$$\int_0^{\tau_k} W_{V_a}(\tau_k - x\mid\lambda_{a_1},\lambda_{a_2})\, g_{U_{ka}}(x)dx \quad\text{and}\quad \int_0^{\tau_l} W_{V_a}(\tau_l - x\mid\lambda_{a_1},\lambda_{a_2})\, g_{U_{la}}(x)dx,$$

respectively. Let $\tau_l = a\tau_k, a\ge 1$, and consider an arbitrary point $(\lambda_{a_1},\lambda_{a_2})\in D$, then

$$\begin{aligned}
\int_0^{\tau_k} W_{V_a}(\tau_k - x\mid\lambda_{a_1},\lambda_{a_2})\, g_{U_{ka}}(x)dx &= \int_0^{a\tau_k} W_{V_a}\left(\tau_k - \frac{u}{a}\mid\lambda_{a_1},\lambda_{a_2}\right)\cdot\frac{1}{a}\cdot g_{U_{ka}}\left(\frac{u}{a}\right)du,\\
&= \int_0^{\tau_l} W_{V_a}\left(\frac{\tau_l - u}{a}\mid\lambda_{a_1},\lambda_{a_2}\right)\hat{g}_{U_{ka}}(u)du,\\
&\le \int_0^{\tau_l} W_{V_a}(\tau_l - u\mid\lambda_{a_1},\lambda_{a_2})\,\hat{g}_{U_{ka}}(u)du,
\end{aligned}$$

where the first equality comes from the substitution $u = ax$, the term $\hat{g}_{U_{ka}}$ in the second equality corresponds to the density function $g_{U_{ka}}$ scaled by a factor $a$. Now, we show that the following inequality holds:

$$\int_0^{\tau_l} W_{V_a}\left(\tau_l - x \mid \lambda_{a_1}, \lambda_{a_2}\right) \hat{g}_{U_{ka}}(x)dx \leq \int_0^{\tau_l} W_{V_a}\left(\tau_k - x \mid \lambda_{a_1}, \lambda_{a_2}\right) g_{U_{la}}(x)dx$$

Consider

$$\begin{aligned}
\int_0^{\tau_l} W_{V_a}\left(\tau_k - x \mid \lambda_{a_1}, \lambda_{a_2}\right) g_{U_{la}}(x)dx - & \int_0^{\tau_l} W_{V_a}\left(\tau_l - x \mid \lambda_{a_1}, \lambda_{a_2}\right) \hat{g}_{U_{ka}}(x)dx, \\
& = \int_0^{\tau_l} W_{V_a}\left(\tau_l - u \mid \lambda_{a_1}, \lambda_{a_2}\right) \left[g_{U_{la}}(u) - \hat{g}_{U_{ka}}\right] du, \\
& = \int_0^{\tau_l} w_{V_a}\left(\tau_l - u \mid \lambda_{a_1}, \lambda_{a_2}\right) \left[G_{U_{la}}(u) - \hat{G}_{U_{ka}}(u)\right] du, \\
& = \int_0^{\tau_l} w_{V_a}\left(\tau_l - u \mid \lambda_{a_1}, \lambda_{a_2}\right) \left[G_{U_{la}}(u) - G_{U_{ka}}\left(\frac{u}{a}\right)\right] du, \\
& \geq 0,
\end{aligned}$$

where the second inequality is the result of applying integration by parts. This allows us to conclude that

$$\int_0^{\tau_k} W_{V_a}\left(\tau_k - x \mid \lambda_{a_1}, \lambda_{a_2}\right) g_{U_{ka}}(x)dx \leq \int_0^{\tau_l} W_{V_a}\left(\tau_l - x \mid \lambda_{a_1}, \lambda_{a_2}\right) g_{U_{la}}(x)dx.$$

Then $S_{T_{ka}}(\tau_k \mid \lambda_{a_1}, \lambda_{a_2}) \leq S_{T_{la}}(\tau_l \mid \lambda_{a_1}, \lambda_{a_2})$, which implies that $f^{ka} \leq f^{la}$ showing that inequality $y_{e(ka)} \leq y_{e(la)}$ is valid. □

## Appendix J: Cutting plane algorithms for models $M2$ and $M3$

Algorithm 3 is the cutting plane algorithm for the solution of model $M2$. Algorithm 4 is the cutting plane method for the solution of model $M3$.

## Appendix K: Accuracy of the homothetic outer approximation

We analyze the accuracy in the computation of the service level when the homothetic approximation $\lambda_{a_1} = \tilde{f}^{ka}(\lambda_{a_2})$ is considered instead of the actual function $\lambda_{a_1} = f^{ka}(\lambda_{a_2})$. The point $\left(\tilde{f}^{ka}(\lambda_{a_2}), \lambda_{a_2}\right)$ can be interpreted as an approximation to the point $(f^{ka}(\lambda_{a_2}), \lambda_{a_2}) \in E_\alpha^{S_{T_{ka}}}$. Hence, $S_{T_{ka}}\left(\tau_k \mid \tilde{f}^{ka}(\lambda_{a_2}), \lambda_{a_2}\right) = \tilde{\alpha}$ is an approximation to the actual service level $\alpha$. To analyze the error $\alpha - \tilde{\alpha}$, we compute $\tilde{\alpha}$ from a sample of 100 points for $\tilde{f}^{ka}(\lambda_{a_2})$ evenly distributed in the interval $\lambda_{a_2} \in [0, \lambda_{ka_2}^{\max}(a_1)]$. The computations are done for all combinations of commodities and hub arcs on the 10 and 20-node instances of the AP and COL datasets, respectively. Figure 14 shows that the average error is of the order of $10^{-3}$, the maximum error is 0.004, and the minimum error is $-1.0 \times 10^{-6}$. The results suggest that the heuristic algorithm 2 is effective in obtaining an accurate outer approximation to the true function.

**Algorithm 3:** Cutting plane generation at integer solutions of the branching three of `M2`

```
Data: $\mathscr{B}, \mathscr{L}_o, \tilde{\varepsilon}, \alpha, (\bar{\mathbf{x}}, \bar{\mathbf{v}}, \bar{\mathbf{z}})$, parameters related to the computation of the total service time
      distribution.
$Cnt \leftarrow 0$;
for $o = (m, n) \in \mathscr{B}$ do
    $CutCnt \leftarrow 0$;
    if $\bar{z}_m = 1$ and $\bar{z}_n = 1$ then
        $\lambda_m \leftarrow \sum_{k \in K} \sum_{\substack{a \in \mathscr{A}_k, \\ m \in a}} w_k \bar{x}_{ka}$;
        $\lambda_n \leftarrow \sum_{k \in K} \sum_{\substack{a \in \mathscr{A}_k, \\ n \in a}} w_k \bar{x}_{ka}$;
        for $l \in \mathscr{L}_o$ do
            if $\bar{v}_l = 1$ then
                if $W_{V_o}(\tilde{\tau}_{lo} \mid \lambda_m, \lambda_n) < \alpha - \tilde{\varepsilon}$ then
                    $(\lambda^{\mathbf{r}}_{lm}, \lambda^{\mathbf{r}}_{ln}) \leftarrow \left[\tilde{f}^{lo}\left(\min\left\{\lambda_n, \tilde{\lambda}^{\max}_{ln}\right\}\right), \min\left\{\lambda_n, \tilde{\lambda}^{\max}_{ln}\right\}\right]$;
                    Add cut (34);
                    $CutCnt \leftarrow CutCnt + 1$;
                end
            end
            if $CutCnt > 0$ then
                $Cnt \leftarrow Cnt + 1$;
            end
        end
    end
end
// For model M3 only.
if $Cnt = 0$ then
    Introduce fine perspective cuts if necessary;
end
```

**Algorithm 4:** Generating fine perspective cuts for model M3 (line 22 of Algorithm 3)

**Data:** $\mathscr{B}, \mathscr{L}_o, \varepsilon, \alpha, (\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$, parameters related to the computation of the total service time distribution.

```
for $o = (m,n) \in \mathscr{B}$ do
    $CutCnt2 \leftarrow 0$;
    if $\bar{z}_m = 1$ and $\bar{z}_n = 1$ then
        $\lambda_m \leftarrow \sum_{k \in K} \sum_{\substack{a \in \mathscr{A}_k, \\ m \in a}} w_k \bar{x}_{ka}$;
        $\lambda_n \leftarrow \sum_{k \in K} \sum_{\substack{a \in \mathscr{A}_k, \\ n \in a}} w_k \bar{x}_{ka}$;
        for $l \in \mathscr{L}_o$ do
            if $\bar{v}_l = 1$ then
                if $\lambda_n > \lambda^{max}_{ln}(m)$ or $\lambda_m > \lambda^{max}_{lm}(n)$ then
                    Add cuts (30);
                    $CutCnt2 \leftarrow CutCnt2 + 1$;
                else if $S_{T_{lo}}(\tau_l \mid \lambda_m, \lambda_n) < \alpha - \varepsilon$ then
                    Add cut (29);
                    $CutCnt2 \leftarrow CutCnt2 + 1$;
                end
            end
            if $CutCnt2 > 0$ then
                Add constraint $x_{ka} \le y_l$, for $l = e(ka)$, if such constraint has not been added
                 previously to the optimization problem;
            end
        end
    end
end
```

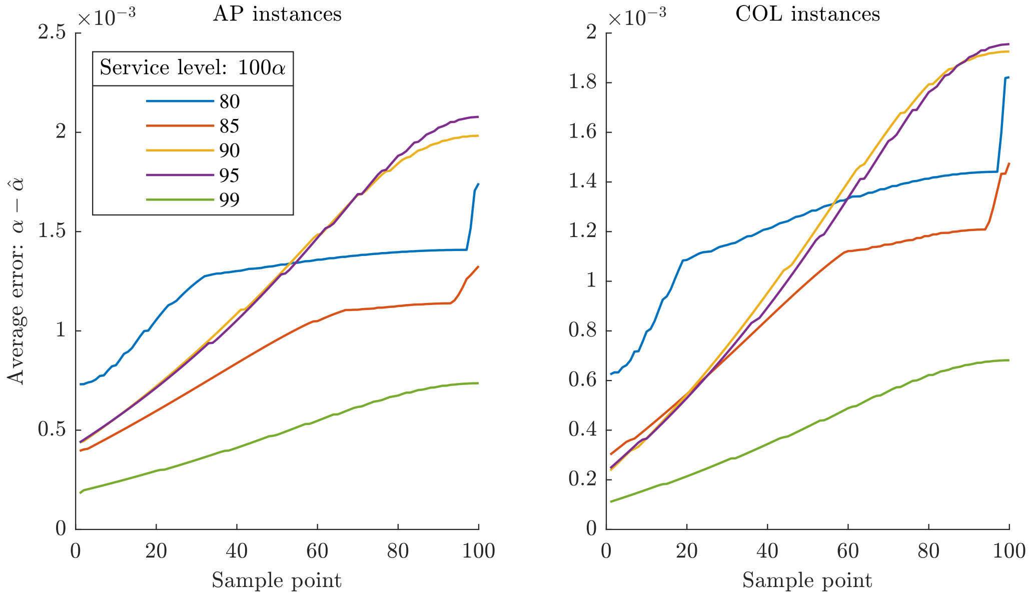


**Figure 14** **Approximation error. Average errors are listed in ascending order.**