
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247223/

Version: Published Version

Article:
Osorio-Mora, A., Álvarez-Miranda, E. and Saldanha-da-Gama, F. (2026) The latency location-or-routing problem. Transportation Research Part E: Logistics and Transportation Review, 216. 105280. ISSN: 1366-5545

https://doi.org/10.1016/j.tre.2026.105280

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.





Transportation Research Part E 216 (2026) 105280

Contents lists available at ScienceDirect

Transportation Research Part E

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

The latency location-or-routing problem

Alan Osorio-Mora [a,*], Eduardo Álvarez-Miranda [b,c,d], Francisco Saldanha-da-Gama [e]

[a] Chair of Logistics, School of Business and Economics, RPTU Kaiserslautern-Landau, Gottlieb-Daimler-Straße 42, 67663, Kaiserslautern, Germany
[b] Department of Industrial Engineering, Facultad de Ingeniería, Universidad de Talca, Sede Curicó, Curicó, Chile
[c] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile
[d] Instituto de Ecología y Biodiversidad (IEB), Chile
[e] Sheffield University Management School, Conduit Road, S10 1FL, Sheffield, UK

ARTICLE INFO

Keywords:
Vehicle routing
Set covering
Location routing
Latency
Choice of service mode

ABSTRACT

This paper introduces the latency location-or-routing problem (LLoRP); a variant of the location-or-routing problem (LoRP) and an extension of the latency location-routing problem. The LLoRP consists of selecting (opening) a set of depots to serve a set of customers, which can be visited as part of routes also to be decided. The maximum number of depots to open is assumed to be exogenous. For the customers lying within a certain coverage range from an open depot, direct service is an alternative. A set of capacitated homogeneous vehicles is available to be assigned to the selected depots as needed. The goal is to minimize the sum of the waiting times of the customers for being served (latency). Several customer-centric applications of the LLoRP are discussed. Three mixed-integer-linear-programming formulations are developed and two lower bounds on the optimal latency value are devised. A tradeoff between LLoRP and LoRP is considered within a bi-objective modeling framework that is tackled using the ε−constrained method. An outcome of this analysis are compact formulations for the standard LoRP. This type of model is in itself a novelty for that problem. Finally, extensive computational results are discussed to assess the relevance of the LLoRP and the efficacy of the models and bounds proposed. The results highlight multiple distinguishing features of the LLoRP while identifying the most promising optimization models to use. They also provide strong evidence for the relevance of looking into tradeoff solutions when a cost-oriented objective becomes of interest in addition to latency.

1. Introduction

The location-or-routing problem (LoRP) is a transportation problem introduced by Arslan (2021) that combines the set covering problem (SCP) with the location-routing problem (LRP). It is a problem in which the customers can either be served by a vehicle within a route or by direct allocation from an open depot when they lie within its coverage range. The objective function accounts for the total cost, which depends on the specific setting analyzed. Nevertheless, without loss of generality, it can be defined as the sum of the fixed costs for opening or selecting depots, and the variable transportation cost (including the direct allocation and the routing). An important assumption of LoRPs is that the cost—or travel time—associated with each edge depends on the service mode selected. For instance, although the distance is the same, traversing an edge by walking takes (in general) a longer time than using a vehicle. However, if one thinks of cost, e.g., associated with fuel consumption or CO_2 emissions, it is usually larger when using vehicles than walking or using a bike (direct allocation).

* Corresponding author.
E-mail addresses: alan.osoriomora@rptu.de (A. Osorio-Mora), ealvarez@utalca.cl (E. Álvarez-Miranda), francisco.saldanha-da-gama@sheffield.ac.uk (F. Saldanha-da-Gama).

https://doi.org/10.1016/j.tre.2026.105280
Received 27 December 2025; Received in revised form 9 July 2026; Accepted 23 September 2026
Available online 2 October 2026
1366-5545/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Open access funding enabled and organized by Projekt DEAL.

1.1. Motivation

LoRPs are inspired by real-life situations, where location decisions, service mode selection, vehicle allocation, and routing decisions must be integrated. Examples include retail e-commerce networks, where customers may either collect their orders at a facility or receive them through vehicle deliveries; school transportation systems, where students living close to a school can walk while others require bus transportation; and food-delivery services that combine in-store pickup and home delivery.

The practical relevance of such systems is reflected in their scale. Worldwide retail e-commerce sales exceeded USD 6 trillion in 2024 (eMarketer, 2024), while parcel-delivery networks process hundreds of billions of shipments annually (Bowes, 2023), increasing the need for efficient last-mile transportation systems that combine pickup and delivery options. Likewise, according to the New York School Bus Contractors Association[1], approximately 26 million students are transported daily by about 480,000 school buses in the United States, generating roughly 10 billion student trips per year. Food-delivery systems constitute another important example, with the global online food-delivery market valued at approximately USD 289 billion in 2024 and expected to exceed USD 500 billion by 2030 (Grand View Research, 2025).

A common characteristic of all these applications is that direct service is only feasible within a limited coverage range. Depending on the context, this range may correspond to a maximum walking distance, a service range of a pickup facility, or the maximum operating range of a drone (when the direct service is provided from the facility). Moreover, the continuous growth of e-commerce and distribution systems has increased customer expectations regarding service responsiveness. Customers increasingly value short lead times and rapid access to goods and services, making waiting times a critical performance measure (Forbes, 2023; World Economic Forum, 2024). Indeed, in some markets, waiting time is a key competitive dimension: survey evidence indicates that about two thirds of consumers stop purchasing from a retailer after two to three late deliveries (Voxware, 2019). Consequently, decision-makers face challenges that go beyond cost minimization.

In vehicle-routing contexts, latency is typically defined as the sum of customer waiting times. Latency-routing problems arise in customer-centric settings—such as those discussed above—as an alternative to traditional cost-based optimization (Ngueveu et al., 2010; Moshref-Javadi and Lee, 2016). For instance, Bruni et al. (2023) study a last-mile parcel-delivery problem in Turin through a latency location-routing (LLRP) framework that accounts for uncertain travel times.

One of the main application domains of latency-based problems is humanitarian logistics, where rapid response is essential for reducing human suffering. For example, Bruni et al. (2020) consider a routing assessment problem motivated by the 2010 Haiti earthquake, although the setting involves a single depot. While these studies aim to improve system responsiveness, they do not simultaneously consider location, routing, and direct-service decisions. A pure latency location-routing problem (or latency-VRP) is therefore insufficient to capture the flexibility often available in practice, where aid and goods can be distributed through a combination of transportation modes. Consider, for example, the 2015 Nepal earthquake, which killed nearly 9000 people and injured approximately 100,000 others (UNWFP, 2016). Indeed, a related report by the United Nations World Food Programme (UNWFP) documents the logistics response and gives relevance to the problem we are investigating in this paper (see UNWFP 2016). That report focuses on the effectiveness of the post-disaster operations. One important component in the decision making process was the establishment of the Kathmandu Humanitarian Staging Area. Moreover, evidence is provided that the regional forward operating hubs in heavily impacted, mountainous terrain like Deurali (Gorkha district) and Chautara (Sindhupalchok district) were fundamental for the success of the operations. The report discusses, for instance, how the centralized regional hubs were used for bulk supply storage and distribution, allowing local populations to access relief materials directly, while specialized air assets (the United Nations Humanitarian Air Service—UNHAS helicopters) were used for delivering to completely cut-off, higher-altitude coordinates. Although this is only one example, it illustrates the relevance of distribution systems that combine facility-location, routing, and direct-service decisions. More broadly, the UNWFP assisted over 120 million people facing severe hunger in 2024 through large-scale humanitarian operations involving distribution centers and transportation resources (UNWFP, 2026). These examples underline the importance of studying location-or-routing systems under latency-based objectives, particularly in humanitarian contexts where responsiveness is often more critical than transportation cost.

1.2. Problem setting

In this paper we introduce the latency location-or-routing problem (LLoRP). The objective is to minimize the sum of the arrival (or waiting) times at the customers, i.e., the total latency. The LLoRP is a variant of the LoRP that also extends the latency location-routing problem (Moshref-Javadi and Lee, 2016) by considering direct allocation as an option for serving customers. Accordingly, we adopt several assumptions, derived from the LLRP, namely:

- There is a maximum number of homogeneous capacitated vehicles that can be used.
- There is a maximum number of uncapacitated depots that can be selected.
- Vehicles are assigned to depots as needed, making fleet allocation across depots part of the decision-making process.
- There are no length constraints for the routes.

In the LoRP, vehicles return to their departure depots (closed routes). In contrast, the return trip is irrelevant in the LLoRP because only customer arrival times are considered, allowing the problem to be modeled with open routes. Although some of these assumptions

[1] NYSBCA School Bus Fast Facts

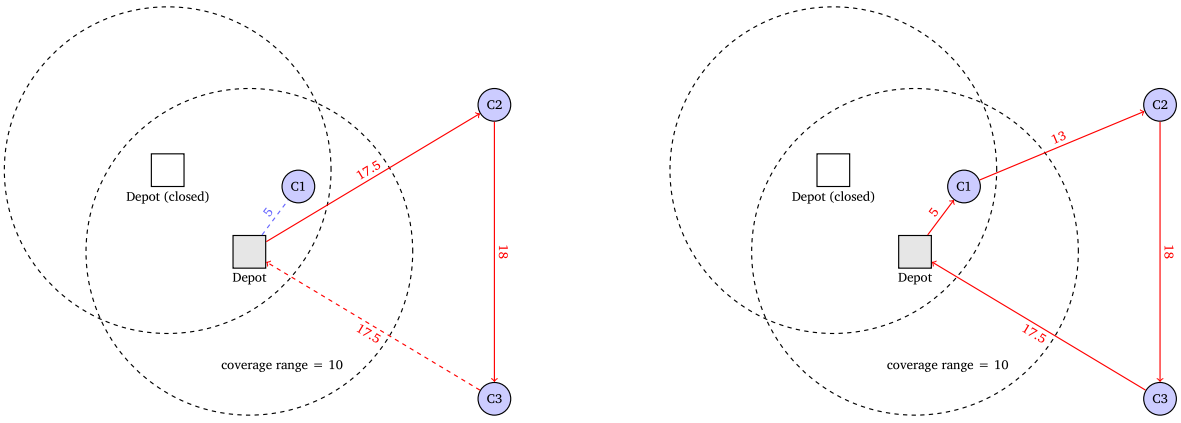


Fig. 1. Optimal solution when considering (**a**) latency (LLoRP), (**b**) cost (LoRP).

differ from those considered in the standard LoRP, there are many features of the LLoRP that are gathered from it (Arslan, 2021). Beyond the similarities and differences between LLoRP and LoRP, the nature of the objective function itself can strongly influence the structure of the optimal solution of the problem, making the LLoRP worth analyzing separately from the LoRP (see, e.g., Sze et al. 2017). In fact, we devote some computational tests to highlight significant differences between both problems under several combinations of assumptions (namely including or neglecting depot-capacity and route-length constraints). In particular, we show that the solutions obtained for both problems can significantly differ. On the other hand, compared to the LLRP, the LLoRP includes a choice of transportation mode which will also be shown to be of relevance for improving the latency even in cases where the travel times for direct allocation are high.

Irrespective of the case (LR or LoR), one must recognize that cost and latency can be of relevance in some applications. For instance, in disaster operations management, cost often proxies the system efficiency while latency proxies the system fairness (see, e.g., Dönmez et al. 2025 for a recent discussion on these aspects). In cases where cost and latency are conflicting, a bi-criteria modeling framework becomes adequate because it allows determining solutions that represent a tradeoff between both objectives. This approach is also considered in this paper. Fig. 1 illustrates such a conflict using a toy example where one depot must be open and there is a single vehicle available. Let us assume that both travel time and cost are equivalent, and can be represented by the Euclidean distance irrespective of the transportation mode. The optimal solution of the LLoRP is to serve C1 with direct allocation, which is a sub-optimal decision for the LoRP. The latency value associated with the optimal LLoRP solution is 58 (arrival times at C1, C2, and C3 are 5, 17.5, and 35.5, respectively), while the total cost is 58 (sum of all the edges in the solution, including the one incident to the depot). On the other hand, the latency associated with the optimal LoRP solution is 59 (arrival times at C1, C2, and C3 are 5, 18, and 36, respectively), while the cost is 53.5. We note that in this example the optimal solution is not unique. For instance, reversing the direction of the arcs in Fig. 1(a) leads to the same cost and latency although the order for visiting the customers changes.

1.3. Contributions

The current work goes beyond the existing literature in multiple aspects, which we can enumerate as follows: (i) our study unifies the existing knowledge for both the LLRP and LoRP. In other words, the latter two problems emerge as particular cases of the problem we are investigating. Technically this can be of relevance because a breakthrough involving the new unified case directly translates into a breakthrough to the particular cases. (ii) Three compact MILP models are introduced for the new problem. (iii) Related to the previous aspect, the newly proposed models are adapted for solving the cost-based LoRP and compared to the state-of-the-art methods (Arslan, 2021; Haghi et al., 2023; Zhou et al., 2026). To the best of our knowledge there is no other compact formulation in the literature for effectively solving the standard LoRP. This may be of major relevance for practitioners who may not master important algorithmic skills within discrete optimization. (iv) Two lower bounds are presented and discussed for the new problem. (v) A bi-objective modeling framework is considered as a tool to show that cost can hardly be taken as a proxy for latency in a location-or-routing setting—instead, those objectives may easily be conflicting thus calling for a tradeoff to be sought when both are of relevance. (vi) An extensive managerial analysis is conducted for the LLoRP. In particular, the impact on the waiting times of the customers from considering a hybrid transportation system (vehicles and direct allocation) instead of a vehicle-based one (LLRP) is highlighted. Moreover, the tradeoff between latency and two design parameters related to the coverage range and direct allocation, is clearly evidenced and analyzed. Furthermore, the effect that the maximum number of depots to open and the capacity of the vehicles have on the waiting times of the customers and on the structure of the solutions is analyzed. Finally, the ϵ-constrained method is used for studying the tradeoff between LLoRP and LoRP within a bi-objective setting. The structural differences between the optimal solutions of both problems are also evidenced.

The remainder of the paper is organized as follows. In Section 2 the related literature is reviewed. In Section 3 the problem is formally described and three optimization models are presented. In Section 4 several insights are provided and two lower bounds

are discussed. In Section 5 the modified formulations for solving the cost-based LoRP are presented and the latency-cost tradeoff is discussed in the context of a bi-objective modeling framework. Section 6 reports on the results of the extensive computational experiments conducted to assess the methodology proposed and insightful managerial insights. The paper ends with an overview of the work done and some suggestions for future research.

2. Literature review

The literature informing our work can be organized into three streams: (i) location-or-routing problems, in which customers within the coverage range of a facility may be served directly instead of being visited by a vehicle (Section 2.1); (ii) latency location-routing problems, in which facility and routing decisions are integrated under a customer-centric objective (Section 2.2); and (iii) routing studies addressing the tradeoff between cost and latency (Section 2.3). Each stream closes with the statement of a specific research gap. At the end of this section Table 1 summarizes the related literature and explains how this paper fills the three gaps. For updated overviews and applications of LRPs, the reader is referred to Schneider and Drexl (2017) (standard LRP), Mara et al. (2021) (different variants of LRPs), Arslan et al. (2021) (humanitarian logistics), Nasir and Kuo (2024) (home healthcare), Wang et al. (2025) (environmentally sustainable routing), and Cavagnini et al. (2025) (recent developments).

2.1. Location-or-routing problems

The LoRP was introduced by Arslan (2021) with different variants and extensions: capacitated and uncapacitated, p-median/p-center, maximal covering, set covering, and hard-cost minimizing. A set partitioning formulation was proposed by the author and solved using a branch-and-price (BP) algorithm. An extension of the LoRP considering a distance-decaying function for the facility coverage and allowing the vehicles to partially cover the customers was investigated by Haghi et al. (2023). The objective is to maximize the number of covered customers. In case of multiple optimal solutions, the percentage of the spent budget is minimized as a secondary objective. The authors proposed two vehicle-dependent three-index formulations for modeling the problem (a compact and a non-compact one). In both cases, it is difficult to solve even tiny instances to proven optimality. Consequently, the problem is handled using an adaptive large neighborhood search (ALNS) heuristic that also provided good results for the standard LoRP. Wolfinger et al. (2023) studied a variant of the LoRP inspired by the need to plan for testing potential COVID-19 cases. There are several differences w.r.t. the base problem. First, the depots where the vehicles are based are different facilities from the test centers. The latter are the facilities that can directly provide the service. Second, the location decision regards the test centers only. Moreover, the vehicles are uncapacitated, and time windows are assumed. The authors modeled the problem using a vehicle-dependent three-index formulation, which they could solve to proven optimality within 24 hours only for instances with 10 customers. Feasible solutions to the problem are sought using a large neighborhood search algorithm. Another variant of the problem with a practical application in forest fire monitoring systems is studied by De la Fuente et al. (2024). In this case, the areas can be monitored using three technologies: watchtowers, balloons and drones. The covering location problem involves the watchtowers and balloons, while the design of the routes is considered for the drones, that are uncapacitated. There is a set of existing watchtowers, and other candidates to be located. There is a given number of drones for each watchtower. Their cover range is limited by the battery autonomy. The authors developed a matheuristic algorithm for finding feasible solutions to the problem. They also present a case-study comprising real-life instances. From a modeling perspective, De la Fuente et al. (2024) proposed two-index formulations (compact and non-compact). The compact one was proved to be effective. Nevertheless, as explained above, that problem has some structural differences w.r.t. the standard LoRP, i.e., their formulations cannot be directly applied to the LoRP. Consequently they did not perform experiments for that problem. Regarding the non-compact formulation, it was merely used for explaining the problem under study (no computational experiments were performed with this method). Finally, Zhou et al. (2026) study an extension of the LoRP inspired in last-mile delivery logistic systems. The referred problem considers different types of customers, namely those willing to be served by robot deliveries (within a time window), those willing to be served by direct allocation (pickup at lockers), and those that accept any of the two mentioned options. Furthermore, the robots can perform multiple trips with a maximum total (accumulated) distance. The authors modeled the problem using a three-index formulation indexed over the vehicles (trips), which is not able to solve even small-sized instances. The solution method is a variable neighborhood search algorithm (called MVNS), which is also tested for the standard LoRP, outperforming the state-of-the art methods when considering a subset of benchmark instances. To the best of our knowledge, no other work has been published that addresses a LoRP.

All LoRP variants studied so far are cost-oriented: a customer-centric (latency) objective has never been considered in a location-or-routing setting, although the applications reviewed above (humanitarian relief, contagious disease testing, last-mile delivery) are precisely those where responsiveness matters most. Moreover, there is no effective compact formulation for the LoRP: the state-of-the-art set-partitioning formulation must be solved by sophisticated column-generation-based methods rather than by an off-the-shelf solver (see Arslan 2021), which is a drawback for practitioners lacking expertise in such sophisticated tools. Moreover, such a type of model is usually harder to adapt to variants and extensions of a problem than compact formulations.

2.2. Latency location-routing problems

The latency location-routing problem (LLRP) is a variant of the LRP that seeks to minimize the latency of the system. It does not involve fixed costs in the objective function. Additionally, the number of vehicles is fixed and the number of depots to open is

limited. The LLRP generalizes the multi-depot cumulative capacitated vehicle routing problem (MDCCVRP) introduced in Lalla-Ruiz and Voß (2020), where all the available depots can be used. Similarly, the MDCCVRP generalizes the cumulative capacitated vehicle routing problem (CCVRP), presented in Ngueveu et al. (2010), where there is only one depot. Although the problems are related to the LLoRP, they do not lie within the scope of the current paper. For the particular case of the CCVRP, the reader is referred to Corona-Gutiérrez et al. (2022) for solution methods and variants.

The LLRP was introduced by Moshref-Javadi and Lee (2016), who introduced a three-index formulation (where the edge variables are indexed over the vehicles) and developed a memetic algorithm as well as a recursive granular algorithm. The most efficient compact formulations for the LLRP, are based on a multi-layer network. Multi-layer formulations (MLF) were introduced by Nucamendi-Guillén et al. (2022), where also a GRASP-based heuristic was proposed. More recently, metaheuristics combining simulated annealing and variable neighborhood descent were proposed by Osorio-Mora et al. (2023b). A multi-start iterated local search algorithm, called M-ILS, was developed by Osorio-Mora et al. (2023a). The state-of-the-art metaheuristic algorithm for solving the LLRP is the hybrid evolutionary algorithm presented in Zou et al. (2024).

Some extensions of the LLRP have been studied in the literature. The drone-LLRP with uncertain flight times was investigated by Bruni et al. (2023). The authors consider a non-linear energy consumption function and cast the problem in a robust optimization context. The LLRP with stochastic travel times was introduced by Osorio-Mora et al. (2025). The problem is cast as a two-stage decision-making problem where the depots to open are first-stage decisions while the routing decisions are adaptive. The LLRP with split deliveries is studied by Nucamendi-Guillén et al. (2026), who devised an MILP formulation and developed an iterated local search algorithm for finding feasible solutions.

To the best of the authors' knowledge, no work in the literature considers an LLRP with direct-allocation service. Consequently, it remains unclear to what extent, and under which conditions, customer waiting times can be reduced by adopting a hybrid transportation system combining routes and direct access to the facilities. From a methodological point of view, the multi-layer formulation—the most effective approach for latency routing problems (Nucamendi-Guillén et al. 2022)—handles open routes only, which prevents its use in settings where each vehicle must return to the depot it departed from; this becomes essential as soon as a cost objective enters the analysis, as in the bi-objective setting studied in this paper.

2.3. The tradeoff between latency and cost

One distinguishing feature of our work is the analysis of the conflicting nature of latency and cost. This aspect has scarcely been analyzed in the literature. A notable exception is due to Sze et al. (2017), who sought a tradeoff between both objectives in the context of the vehicle routing problem (i.e., no depot selection, and no direct allocation allowed). The authors compared upper bounds obtained via an heuristic algorithm. A similar case is presented in Molina et al. (2018) where the tradeoff between travel cost, maximum latency and fleet size is studied. A multi-start heuristic algorithm is proposed. Recently, Segredo et al. (2026) studied a bi-objective (cost versus latency) MDCCVRP. The authors considered open routes under both objective functions. A matheuristic framework combining the well-known non-dominated sorting genetic algorithm II (NSGA-II) and the augmented ε-constraint method was proposed as the solution approach. Within the ε-constraint framework, the authors employed a standard MLF. To the best of the authors' knowledge, there is no other work accounting for the tradeoff between cost and latency in routing problems. The latency–cost tradeoff has so far been explored mainly by comparing heuristic upper bounds, and only in pure routing settings—without location decisions and service-mode choices. An exact bi-objective analysis of an integrated location-or-routing problem, preserving the closed-route structure of the cost-based problem, is missing.

This paper addresses the three gaps jointly. The first gap is filled by introducing the LLoRP and by proposing three compact formulations that, if suitably adapted (Section 5.1), also solve the standard cost-based LoRP with an off-the-shelf solver, providing new proven optimal solutions and improved upper bounds for benchmark instances (Section 6.2). The second gap is filled by embedding the direct-allocation option in a latency location-routing setting and quantifying the waiting-time savings it enables (Section 6.3.1). Additionally, a new multi-layer formulation with closed routes is introduced (see Sections 3.3 and 5.1). Finally, the third gap is filled by the exact bi-objective analysis of Sections 5 and 6.3.4, which assesses the conflict between cost and latency at all three decision levels (location, fleet, and routing) and under different assumptions. This analysis is based on formulations with closed routes.

To better highlight the contribution provided by the current paper, we summarize the reviewed literature in Table 1. The first column in this table details the reference. Columns headed by 'L', 'R', and 'DA' specify whether Location, Routing or Direct Allocation decisions are made. We add some observations on the specific problem analyzed in each source (column 'Observation'). The column "Objective function" specifies the performance measure(s) considered in each reviewed paper whereas the following column, headed "Bi-objective" indicates whether the tradeoff between cost and latency is analyzed in the source. Finally, column "Formulation" details the type of mathematical model used for modeling the analyzed problem. The notation of this column is the following: •3-IF: Compact flow-based formulations where the edge variables are indexed over other depots; •3-IF-V: Compact formulations where the edge variables are indexed over the vehicles; •NC3-IF-V: Non-compact formulations where the edge variables are indexed over the vehicles and the subtour elimination constraints are exponential in size (usually solved with branch-and-cut algorithms); •SPF: Non-compact set partitioning formulation formulations (usually solved with column-generation based algorithms); •MLF: Compact multi-layer formulations. These formulations consider open routes (edges incident to the depot are not considered); •MLFC: Compact multi-layer formulations with closed routes; •2-IF: Compact flow-based formulations where the edge variables are not indexed over other vehicles or depots; •NC2-IF: Non-compact formulations where the edge variables are not indexed over other vehicles or depots and the subtour elimination constraints are exponential in size (usually solved with branch-and-cut algorithms).

Table 1
Summary of the related literature.

Authors	L	R	DA	Observation	Objective Function	Bi-objective	Formulation
(Arslan, 2021)	✓	✓	✓	standard LoRP.	Cost		SPF
(Bruni et al., 2023)	✓	✓		LLRP variant with drones and uncertain travel times.	Latency		MLF
(De la Fuente et al., 2024)	✓	✓	✓	LoRP variant with uncapacitated vehicles (drones), no customers demands, and existing facilities.	Cost		NC2-IF, 2-IF
(Haghi et al., 2023)	✓	✓	✓	LoRP variant with decaying coverage; standard LoRP.	Cost (modified)		NC3-IF-V, 3-IF-V
(Lalla-Ruiz and Voß, 2020)		✓		standard MDCCVRP.	Latency		3-IF-V
(Molina et al., 2018)		✓		CCVRP variant with multi-trips.	Latency, Cost, Fleet size	✓	
(Moshref-Javadi and Lee, 2016)	✓	✓		standard LLRP.	Latency		3-IF-V
(Ngueveu et al., 2010)		✓		standard CCVRP.	Latency		3-IF-V
(Nucamendi-Guillén et al., 2022)	✓	✓		standard LLRP.	Latency		MLF
(Nucamendi-Guillén et al., 2026)	✓	✓		LLRP variant with split deliveries.	Latency		MLF
(Osorio-Mora et al., 2025)	✓	✓		LLRP variant with stochastic travel times.	Latency		MLF
(Segredo et al., 2026)		✓		MDCCVRP.	Latency, Cost	✓	MLF
(Sze et al., 2017)		✓		standard CCVRP.	Latency, Cost	✓	
(Wolfinger et al., 2023)	✓	✓	✓	LoRP variant with time windows, uncapacitated vehicles and two types of facilities.	Cost		3-IF-V
(Zhou et al., 2026)	✓	✓	✓	LoRP variant with time windows, different types of customer preferences and multiple-trips; standard LoRP.	Cost		3-IF-V
This paper	✓	✓	✓	LLoRP; standard LoRP.	Latency, Cost	✓	2-IF, 3-IF, MLF, MLFC

3. Problem description and mathematical formulations

To better describe the LLoRP and formulate it mathematically, we start by introducing some related notation and a few additional assumptions. We are combining aspects from the LLRP (Moshref-Javadi and Lee 2016) with others from the LoRP (Arslan 2021). We consider the following sets: C is the set of customers, D is the set of potential locations for the depots (assumed to be uncapacitated), V is the set of nodes, $V = C \cup D$, and K the set of homogeneous capacitated vehicles in the available fleet. Different parameters are involved in the problem. p denotes the maximum number of depots that can be selected; R is the coverage range of the depots e.g., guided by the perceived acceptable walking time; d_i represents the demand of customer $i \in C$, with $d_i \geq 0$; t_{ij} is the travel time between nodes i and j ($i, j \in V, i \neq j$, with $t_{ij} = t_{ji}$); Q is the capacity of the vehicles.

The LLoRP consists of selecting at most p depots to open and determining the routes or sequence of visits to make for at most $|K|$ homogeneous capacitated vehicles. Every customer must be served once either by a depot (direct allocation) or by a vehicle (within a route).

Like in the standard LLRP and LoRP, we assume that splitting the demands is not allowed. The objective is to minimize the latency of the system, i.e., the sum of the waiting times of the customers.

Like in the cost-oriented LoRP, we must balance direct-allocation and routing costs in the objective function (see Arslan 2021). If this is not done, one of those components may become dominant thus reducing the problem to a pure location-allocation or location-routing problem. With this purpose, two cost elements are introduced by Arslan (2021) that we also consider in our work: F_R and F_A denote multiplication factors to consider for computing the travel times when the customers are served within a route and directly by a depot, respectively.

Last but not least, an important element when modeling a latency location-routing problem is an upper bound on the number of nodes in an open route. We denote it by M. It can be easily computed by sorting the demands of the customers non-decreasingly and counting the number of customers whose accumulated demand does not exceed Q. M is equal to that maximum plus one (the depot).

Before presenting the formulations, we pause on the assumptions of the LLoRP and on their practical justification. (i) The latency objective (sum of the arrival times) reflects deprivation in post-disaster distribution—where each hour before relief reaches a household prolongs suffering (UNWFP 2016; Dönmez et al. 2025)—and customer waiting in commercial delivery, where retention is highly sensitive to delivery delays (Voxware, 2019). (ii) Imposing a maximum number p of depots, exogenously fixed, mirrors the practice of activating relief facilities among existing infrastructures (schools, stadiums, parking lots) under budget and staffing caps, as done with the forward hubs in the Nepal operation (UNWFP 2016); in commercial settings, p reflects the number of stores or micro-depots a chain can dedicate to fulfillment. Since such facilities are pre-existing, no fixed opening cost enters the base model, and a cap on their number acts as a budget constraint with identical fixed costs. (iii) Depots are uncapacitated because staging areas and stores hold bulk stock that is replenished independently of the daily distribution flow; the Kathmandu Humanitarian Staging Area, used for bulk supply storage and distribution, is an example (UNWFP 2016). The impact of explicitly adding such constraints is nonetheless quantified in Section 6.3.4. (iv) The fleet is homogeneous and capacitated, as humanitarian fleets and last-mile operators typically

standardize on one vehicle type (e.g., identical trucks, vans, or cargo bikes), whose payload defines Q; the derived bound M on the number of nodes per route may equivalently stem from technical restrictions such as energy autonomy (Bruni et al. 2023). (v) No route-length constraints are imposed: in the urban and regional operations motivating the LLoRP, route durations are well below shift limits; the impact of explicitly adding such constraints is nonetheless quantified in Section 6.3.4. (vi) The coverage range R represents the perceived acceptable effort for reaching a facility. This has direct counterparts in practice: humanitarian standards prescribe maximum walking distances to distribution and water points (e.g., at most 500 m to the nearest water point according to Sphere Association 2018), while in retail it captures the distance a customer accepts to travel for in-store or locker pickup; when the direct mode is operated by drones, R is the flight range (Bruni et al. 2023). (vii) The factors F_R and F_A convert one distance into mode-dependent travel times: walking at 4–5 km/h against urban driving at 25–40 km/h yields F_A values severalfold larger than F_R, which is the range explored in our experiments ($F_A/F_R \in [1,6]$); the converse case $F_A < F_R$ arises when the direct mode is faster point-to-point technology such as drones, nevertheless in such cases the direct allocation will always be preferred over routing (see Section 4). (viii) Demands are not split, as each household receives its relief kit (or each customer their order) in a single service. (ix) Service times are not modeled explicitly: when they are approximately constant across customers, they can be absorbed into the travel times of the outgoing edges without altering the structure of the problem.

3.1. A two-index formulation

A two-indexed formulation that we denote by (2-IF) can be proposed for the LLoRP considering the following sets of decisions variables: z_j is equal to 1 if depot $j \in D$ is selected and 0 otherwise; s_{ij} is equal to 1 if customer $i \in C$ is directly served by depot $j \in D$ and 0 otherwise; r_{ij} is equal to 1 if the edge going from node i to node j ($\forall i, j \in V : i \neq j$) is in the solution, and 0 otherwise; w_{ij} is a non-negative variable representing the load traversing the edge going from node i to node j ($\forall i, j \in V : i \neq j$). Finally, n_{ij} represents the number of customers affected by the edge going from node i to node j ($\forall i, j \in V : i \neq j$). In other words, this is the number of remaining customers after visiting node i if the vehicle visits j next in the route. Notice that in routing problems, the latency can be computed as the product between the travel time associated with each edge in the solution and the number of remaining customers in the route after traversing the edge. For example, consider a route starting from the depot (node 0) that visits three customers, say j_1, j_2, j_3 in this order. The latency of the route can be computed as $3t_{0j_1} + 2t_{j_1j_2} + t_{j_2j_3}$, which is equivalent to the sum of the arrival times at each customer: arrival time at customer j_1 is t_{0j_1}, at customer j_2 is $t_{0j_1} + t_{j_1j_2}$, and at customer j_3 is $t_{0j_1} + t_{j_1j_2} + t_{j_2j_3}$.

The LLoRP can be formulated mathematically as follows:

minimize
$$F_R \sum_{i \in V} \sum_{j \in V \setminus \{i\}} t_{ij} n_{ij} + F_A \sum_{i \in C} \sum_{j \in D} t_{ij} s_{ij}, \tag{1}$$

subject to
$$\sum_{j \in D} z_j \leq p, \tag{2}$$
$$\sum_{i \in D} \sum_{j \in C} r_{ij} \leq |K|, \tag{3}$$
$$\sum_{j \in V \setminus \{i\}} r_{ji} + \sum_{j \in D} s_{ij} = 1 \qquad \forall i \in C, \tag{4}$$
$$\sum_{j \in V \setminus \{i\}} r_{ij} + \sum_{j \in D} s_{ij} = 1 \qquad \forall i \in C, \tag{5}$$
$$r_{ji} \leq z_j \qquad \forall i \in C, j \in D, \tag{6}$$
$$r_{ij} \leq z_j \qquad \forall i \in C, j \in D, \tag{7}$$
$$\sum_{j \in V \setminus \{i\}} r_{ij} = \sum_{j \in V \setminus \{i\}} r_{ji} \qquad \forall i \in V, \tag{8}$$
$$R\, z_j \geq t_{ij}\, s_{ij} \qquad \forall i \in C,\ j \in D, \tag{9}$$
$$z_j \geq s_{ij} \qquad \forall i \in C,\ j \in D, \tag{10}$$
$$\sum_{j \in V \setminus \{i\}} w_{ji} - \sum_{j \in V \setminus \{i\}} w_{ij} = d_i (1 - \sum_{j \in D} s_{ij}) \qquad \forall i \in C, \tag{11}$$
$$\sum_{j \in V \setminus \{i\}} n_{ji} - \sum_{j \in V \setminus \{i\}} n_{ij} = 1 - \sum_{j \in D} s_{ij} \qquad \forall i \in C, \tag{12}$$
$$w_{ij} \leq Q\, r_{ij} \qquad \forall i, j \in V,\ i \neq j, \tag{13}$$
$$n_{ij} \leq (M-1) r_{ij} \qquad \forall i, j \in V,\ i \neq j, \tag{14}$$
$$w_{ji} \geq d_i\, r_{ji} \qquad \forall i \in C,\ j \in V,\ i \neq j, \tag{15}$$
$$n_{ji} \geq r_{ji} \qquad \forall i \in C,\ j \in V,\ i \neq j, \tag{16}$$
$$r_{ji} + r_{ij} \leq 1 \qquad \forall i, j \in C,\ i \neq j, \tag{17}$$
$$z_j \in \{0,1\} \qquad \forall j \in D, \tag{18}$$
$$s_{ij} \in \{0,1\} \qquad \forall i \in C,\ j \in D, \tag{19}$$
$$w_{ij} \geq 0 \qquad \forall i, j \in V,\ i \neq j, \tag{20}$$

$$n_{ij} \geq 0 \qquad \forall i, j \in V,\ i \neq j, \tag{21}$$

$$r_{ij} \in \{0,1\} \qquad \forall i, j \in V,\ i \neq j. \tag{22}$$

The objective function (1) accounts for the latency of the system. The first expression computes the latency of the routes, while the second term represents the waiting time of the customers when a direct service is provided. Constraint (2) states that at most p depots can be opened. Constraint (3) ensures that at most $|K|$ vehicles are used. Equalities (4) and (5) ensure that each customer is either visited once (exactly by one vehicle) or directly served by one depot. Inequalities (6) and (7) ensure that the vehicles use open depots. Constraints (8) guarantee the flow balancing. Constraints (9) establish that customers can be directly served only by open depots and only if they are within the coverage range of the depot. Inequalities (10) prevent the allocation of customers to closed depots when $t_{ij} = 0$. Note that this can happen for instance if a demand node is located next to a potential location for a depot. Equalities (11) compute the cumulative demand, i.e., the sum of the demands of the remaining customers in the routes. Similarly, Constraints (12) compute the number of remaining customers in the routes. The capacity of the vehicles is ensured in (13). An upper bound for the maximum number of customers in the routes is established by (14). Inequalities (15) define lower bounds for the load in each edge. Constraints (16) state that if an edge is in the solution, then the number of remaining customers must be greater than or equal to 1. Symmetry-breaking valid inequalities are presented in (17). Finally (18)–(22) state the domain of the decision variables. Note that we can set equal to zero (or eliminate from the model) all variables r_{ij}, w_{ij}, and n_{ij} such that $i, j \in D$.

3.2. A three-index formulation

A three-index formulation (3-IF) for the LLoRP uses the following additional binary variables: r'^{k}_{ij} is equal to 1 if and only if the edge going from node i to node j ($\forall i, j \in V,\ i \neq j$) is traversed by a vehicle allocated to depot $k \in D$; f_{ij} is equal to 1 if and only if customer $i \in C$ is served by a vehicle allocated to depot $j \in D$. The new model (3-IF) is the following:

minimize (1),

subject to (2), (9) – (12), (18) – (21),

$$\sum_{j \in D} (f_{ij} + s_{ij}) = 1 \qquad \forall i \in C, \tag{23}$$

$$f_{ij} \leq z_j \qquad \forall i \in C, j \in D, \tag{24}$$

$$\sum_{i \in D} \sum_{j \in C} r'^{i}_{ij} \leq |K|, \tag{25}$$

$$\sum_{j \in C} r'^{i}_{ij} = \sum_{j \in C} r'^{i}_{ji} \qquad \forall i \in D, \tag{26}$$

$$\sum_{k \in V \setminus \{i\}} r'^{j}_{ik} = f_{ij} \qquad \forall i \in C, j \in D, \tag{27}$$

$$\sum_{k \in V \setminus \{i\}} r'^{j}_{ki} = f_{ij} \qquad \forall i \in C, j \in D, \tag{28}$$

$$w_{ij} \leq Q \sum_{k \in D} r'^{k}_{ij} \qquad \forall i, j \in V,\ i \neq j, \tag{29}$$

$$n_{ij} \leq (M-1) \sum_{k \in D} r'^{k}_{ij} \qquad \forall i, j \in V,\ i \neq j, \tag{30}$$

$$w_{ji} \geq d_i \sum_{k \in D} r'^{k}_{ji} \qquad \forall i \in C, j \in V,\ i \neq j, \tag{31}$$

$$n_{ji} \geq \sum_{k \in D} r'^{k}_{ji} \qquad \forall i \in C, j \in V,\ i \neq j, \tag{32}$$

$$r'^{k}_{ji} + r'^{k}_{ij} \leq f_{ik} \qquad \forall k \in D,\ i, j \in C,\ i \neq j, \tag{33}$$

$$\sum_{k \in D} (r'^{k}_{ji} + r'^{k}_{ij}) \leq \sum_{k \in D} f_{ik} \qquad \forall i, j \in C,\ i \neq j, \tag{34}$$

$$r'^{k}_{ij} \in \{0,1\} \qquad \forall k \in D,\ i, j \in V,\ i \neq j, \tag{35}$$

$$f_{ij} \in \{0,1\} \qquad \forall i \in C,\ j \in D. \tag{36}$$

The objective function for (3-IF) is identical to that of (2-IF). Constraints (23) ensure that each customer is allocated to exactly one depot and served either by a vehicle or by direct allocation. Constraints (24) state that customers can only be served by vehicles departing from open depots. Constraint (25) establishes the maximum number of vehicles to be deployed. Constraints (26) ensure that the number of vehicles departed and retrieved is the same at each depot. Constraints (27) and (28) establish that if customer i is assigned to a vehicle allocated to depot j, there must be one ingoing and one outgoing edge to and from i both associated with the same depot j. Constraints (29) are the capacity constraints of the vehicles. Constraints (30) present an upper bound for the maximum number of customers in the routes. Constraints (31) and (32) establish lower bounds for the load and the number of remaining customers on each edge, respectively. The symmetry-breaking valid inequalities are defined in (33) and (34). Finally,

Constraints (35) and (36) define the domain of the new variables. Note that the variables r'^{k}_{ij} and w_{ij} can be set to zero—or simply removed from the model—when $i, j \in D$. Additionally, variables r'^{k}_{ij} can be set to zero or removed from the model if $i, k \in D$ with $i \neq k$, or if $j, k \in D$ with $j \neq k$.

3.3. The multi-layer formulation

Next we present a third optimization model, which is based on a multi-layer network (see also Nucamendi-Guillén et al. 2022). We denote by $L = \{1, 2, \ldots, M\}$ the set of layers in the network. Recall that M is an upper bound on the number of nodes in a route. A layer in the network should be understood in the opposite way to the position of the nodes in a route. In other words, the first layer (second layer, *et cetera*) in the network contains all customers that are the last (second-last, *et cetera*) ones in their respective routes. The latency of a route can be computed as the sum of the values we obtain when multiplying the number of remaining customers in the route after each edge by the travel time associated with the edge. The motivation for considering the multi-layer network is to simplify the way we calculate the number of times each edge impacts the objective function: this is directly given by the level at which the edge appears in the network.

The following sets of decision variables are introduced: $x_{i\ell}$ is equal to 1 if customer $i \in C$ appears at level $\ell \in L$ of the multi-layer network, 0 otherwise; $y_{ij\ell}$ is equal to 1 if the edge (i, j) starting from node $i \in V$, $i \neq j$, incident to customer $j \in C$ that appears at level $\ell \in L$ ($\ell \neq M$) of the multi-layer network, is in the solution; 0 otherwise. When $y_{ij\ell}$ is equal to 1, ℓ represents the number of customers in the route after edge (i, j), including customer j. v_{ij} is a non-negative quantity representing the load of the vehicle when it departs from depot $i \in D$ to visit customer $j \in C$. Finally, u_{ij} is also non-negative and represents the load of a vehicle when it departs from customer $i \in C$ to arrive at customer $j \in C$, $i \neq j$. Additionally, we keep using the s- and z-variables defined in the context of model (2-IF).

The multi-layer formulation that we denote by (MLF) is the following:

$$\text{minimize} \quad F_R \sum_{i \in V} \sum_{j \in C \setminus \{i\}} \sum_{\ell \in L \setminus \{M\}} \ell \, t_{ij} y_{ij\ell} + F_A \sum_{i \in C} \sum_{j \in D} t_{ij} s_{ij}, \tag{37}$$

$$\text{subject to} \quad (2), (9), (10), (18), (19),$$

$$\sum_{\ell \in L} x_{i\ell} + \sum_{j \in D} s_{ij} = 1 \qquad \forall i \in C, \tag{38}$$

$$\sum_{\ell \in L \setminus \{M\}} \sum_{i \in D} \sum_{j \in C} y_{ij\ell} \leq |K|, \tag{39}$$

$$\sum_{i \in C} x_{i1} = \sum_{\ell \in L \setminus \{M\}} \sum_{i \in D} \sum_{j \in C} y_{ij\ell}, \tag{40}$$

$$\sum_{j \in C \setminus \{i\}} y_{ij\ell} = x_{i\ell+1} \qquad \forall i \in C, \ \ell \in L \setminus \{M\}, \tag{41}$$

$$\sum_{i \in V \setminus \{j\}} y_{ij\ell} = x_{j\ell} \qquad \forall j \in C, \ \ell \in L \setminus \{M\}, \tag{42}$$

$$z_i \geq \sum_{\ell \in L \setminus \{M\}} y_{ij\ell} \qquad \forall i \in D, \ j \in C, \tag{43}$$

$$\sum_{h \in D} v_{hi} + \sum_{j \in C \setminus \{i\}} u_{ji} - \sum_{j \in C \setminus \{i\}} u_{ij} = d_i (1 - \sum_{j \in D} s_{ij}) \qquad \forall i \in C, \tag{44}$$

$$v_{ij} \geq d_j \sum_{\ell \in L \setminus \{M\}} y_{ij\ell} \qquad \forall i \in D, \ j \in C, \tag{45}$$

$$u_{ij} \geq d_j \sum_{\ell \in L \setminus \{M\}} y_{ij\ell} \qquad \forall i, j \in C, \ i \neq j, \tag{46}$$

$$v_{ij} \leq Q \sum_{\ell \in L \setminus \{M\}} y_{ij\ell} \qquad \forall i \in D, \ j \in C, \tag{47}$$

$$u_{ij} \leq (Q - d_i) \sum_{\ell \in L \setminus \{M\}} y_{ij\ell} \qquad \forall i, j \in C, \ i \neq j, \tag{48}$$

$$\sum_{\ell \in L \setminus \{M\}} (y_{ij\ell} + y_{ji\ell}) \leq 1 \qquad \forall i, j \in C, \ i \neq j, \tag{49}$$

$$x_{i\ell} \in \{0, 1\} \qquad \forall i \in C, \ \ell \in L, \tag{50}$$

$$y_{ij\ell} \in \{0, 1\} \qquad \forall i \in V, \ j \in C, \ i \neq j, \ \forall \ell \in L \setminus \{M\}, \tag{51}$$

$$v_{ij} \geq 0 \qquad \forall i \in D, \ j \in C, \tag{52}$$

$$u_{ij} \geq 0 \qquad \forall i, j \in C, \ i \neq j. \tag{53}$$

The objective function (37) accounts for the latency of the system through the new variables. Constraints (38) ensure that each customer is served exactly once, i.e., each customer appears in one and only one level of the multi-layer network or is served directly

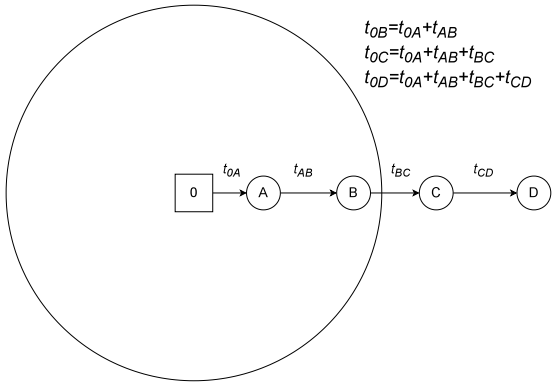


Fig. 2. Particular case when $F_A = F_R$.

by a depot. Inequalities (39) establish the maximum number of vehicles to be used. Constraints (40)–(42) define the sequence in which the nodes appear in the multi-layer network (consequently in the routes) by linking variables x and y. Equalities (40) state that the number of customers visited at the end of each route (level 1 in the multi-layer network) is equal to the number of vehicles performing routes. Constraints (41) ensure that the edge (i, j) is in the solution only if i is visited exactly before j, i.e., if i appears in the immediately upper level in the multi-layer network w.r.t. j (which appears at level ℓ). Constraints (42) guarantee that if a customer j appears at level ℓ, then there must be an incident edge to j. Inequalities (43) impose that the routes start only at open depots. The load of each vehicle at each customer is calculated in Constraints (44). In (45), the load with which each vehicle starts its route is computed. Similarly, Inequalities (46) account for the load with which the vehicle departs from a certain customer, ensuring that it must be at least equal to the demand of the forthcoming customer in the route. The vehicle capacity constraints are ensured by Constraints (47) and (48) for the cases when the vehicle is departing from the depot, or from a customer, respectively. Symmetry-breaking valid inequalities are presented in (49). Finally, (50)–(53) state the domain of the new decision variables.

4. Further insights on the problem and modeling frameworks

One important property of latency routing problems (with vehicles as the only service alternative) is that in an optimal solution, all the available vehicles are used if the number of vehicles is smaller than or equal to the number of customers. This property can be extended to the LLoRP under certain conditions, which will be discussed below. Due to this property, even if the coverage range increases, all the vehicles will be used since it reduces the latency of the system. This is not the case for the cost-based LoRP, where the number of vehicles needed may decrease when the coverage range increases (see Arslan 2021).

The above all-vehicle-usage property is valid if the number of customers outside the covering area of the selected depots is greater than or equal to $|K|$, irrespective of the values of F_A and F_R.

Suppose now that the number of customers outside the covering area of the selected depots is smaller than $|K|$. Three cases emerge:

1. $F_A \leq F_R$.
   In this case, the customers within the range of the open depots will always be served directly by their respective nearest open depot, leading to the use of fewer than $|K|$ vehicles.
2. $F_A > F_R$.
   The customers within the coverage range of the open depots will not necessarily be served by direct allocation. This case is a key part of our experimental analysis (see Section 6.3).
3. $F_A = F_R$ and the optimal solution contains routes with a single customer.
   In this case, multiple optimal solutions exist since visiting the customer with a vehicle is equivalent to the direct allocation.

Another particular situation may occur when $F_A = F_R$. Consider two customers $j, k \in C$ and one depot $i \in D$, with $t_{ij} \leq R$ and $t_{jk} > R$. If $t_{ij} + t_{jk} = t_{ik}$, i.e., if i, j and k lie on the same line in the plane, the waiting times of customers i and j are exactly the same when i is served by direct allocation and when i and j are served in a route with j being visited immediately after i (which is the first customer in the route). This property can be generalized to more than two customers in a line (for Euclidean distances), and is valid since the problem does not consider service times. This particular case is illustrated in Fig. 2, where node 0 is the depot, and A, B, C, and D are the customers. The large circle limits the depot's coverage range. In the solution depicted in the figure, the arrival times at the customers will be the same, irrespective of having A and B served as part of the route or by direct allocation.

Noticeably, if R is set equal to zero, then the problem might be equivalent to the LLRP. This is always true if $t_{ij} > 0, \forall i \in C, j \in D$. We found that for some particular instances when the distance from a certain customer to an open depot is equal to 0, the direct allocation is possible, leading the vehicles to perform more efficient routes due to capacity release (obtaining latency values smaller w.r.t. the equivalent LLRP). Thus, strictly speaking, if direct allocation is not allowed, the problem reduces to the LLRP. Furthermore,

if a solution with p depots such that all customers can be covered by at least one of the open depots exists, and $F_A \leq F_R$, the problem is equivalent to a set covering problem.

Overall, in our problem, the number of vehicles to use ends up being fixed due to the nature of the latency-oriented objective. That number is a decision in the LoRP.

4.1. Lower bounds

Given the properties of the problem discussed above, it is possible to propose two valid lower bounds for its optimal solution value.

The first lower bound is denoted by LB1 and generalizes the lower bound 1 proposed by Moshref-Javadi and Lee (2016) for the LLRP. It stems from relaxing the assumption on the number of vehicles, allowing each customer to be served either by exactly one vehicle (vehicles have a single customer in their routes) or by a direct allocation to an open depot. Let us define an auxiliary variable, say σ_{ij}, equal to 1 if customer $i \in C$ is visited by a vehicle starting from depot $j \in D$ and 0 otherwise. LB1 is the optimal value of the following integer programming problem:

$$\text{minimize} \quad \sum_{i \in C} \sum_{j \in D} t_{ij}(F_A s_{ij} + F_R \sigma_{ij}), \tag{54}$$

$$\text{subject to} \quad (2), (9), (10), (18), (19),$$

$$\sum_{j \in D} (s_{ij} + \sigma_{ij}) = 1 \qquad \forall i \in C, \tag{55}$$

$$\sigma_{ij} \leq z_j \qquad \forall i \in C,\ j \in D, \tag{56}$$

$$\sigma_{ij} \in \{0, 1\} \qquad \forall i \in C,\ j \in D. \tag{57}$$

The objective function (54) minimizes the arrival times at the customers, associated either with vehicles visiting a single customer or by direct allocation. Equalities (55) state that each customer must be served either by a vehicle or by direct allocation. Constraints (56) indicate that the vehicles can start only from open depots. Finally, in (57) the nature of the new variables is presented.

Remark 1. When $F_A \geq F_R$, the solution of the IP presented is equivalent to the one obtained with the lower bound 1 presented in Moshref-Javadi and Lee (2016).

Remark 2. The above problem has the p-median problem as a particular case (when we set $R = 0$ and only variables z and σ remain to be considered). Therefore, it is NP-hard (see, e.g., Laporte et al. 2019 and the references therein). Nevertheless, our results show that the problem can be tackled quite efficiently for instances of the LLoRP that were considered in the experiments whose results are reported in Section 6.

A second lower bound for the LLoRP that we denote by LB2 is obtained by relaxing the capacity constraints of the vehicles and the location decisions. In other words, it stems from considering an MDk-TRP with coverage range for the depots. The idea of using the multi-depot $k-$traveling repairman problem (MDk-TRP) as a lower bound for the LLRP was first proposed by Osorio-Mora et al. (2023b), where it performed as the tightest lower bound for most of the instances studied.

The implementation of LB2 can be done based on the first formulation presented in Bruni et al. (2022) for the MDk-TRP, by adding constraints (19), (38), as well as

$$R\, z_j \geq t_{ij} s_{ij}, \quad \forall j \in D,\ i \in C, \tag{58}$$

and also by replacing the equality of the third constraint presented in Bruni et al. (2022) with an inequality ($\leq |K|$).

In addition to the above lower bounds, in Section 6 we also consider the lower bound provided by the linear relaxation of models (2-IF), (3-IF) and (MLF).

5. Cost versus latency

One relevant research question regards the structural differences (if any exist) between solutions for the LLoRP and those for the LoRP introduced by Arslan (2021). Moreover, despite the customer-centric perspective cast by a latency-oriented objective function, we may face an application in which cost efficiency cannot be disregarded. For instance, humanitarian logistics is prone to settings in which a cost-efficient solution is sought once an acceptable service level is achieved (see, e.g., Bayram et al. 2023).

Seeking to capture these aspects, we propose a vectorial optimization modeling framework considering two objective functions: cost and latency.

We recall that the LoRP seeks to minimize the total cost, which includes the setup costs for the depots and the transportation costs. Regarding the former, we denote by ρ_i the fixed cost for setting up a depot at $i \in D$. In turn, the total transportation cost is computed as the sum of the cost of all the edges in the solution, including those from the last customer in each route that are incident to the depots (that are ignored when we compute the latency). The cost of each edge is assumed to be equal to the product of its travel time and the corresponding cost factor F_R or F_A (depending on whether the customer is served by a vehicle or directly by a depot). We also note that the LoRP includes maximum length constraints for the routes as well as capacities for the depots. We denote by T the length limit and by κ_i the capacity of a depot operating at $i \in D$.

It is important to note that in this paper, F_R and F_A are multiplication factors for the travel time of each edge and not for a cost. Nevertheless, to perform a fair comparison between solutions for the LLoRP and LoRP, we assume that cost is proxied by travel time.

This section is divided into two parts. First, we adapt the already introduced compact formulations to solve the cost-oriented LoRP (Arslan, 2021). Afterwards, we discuss the bi-objective modeling framework and the methodology for determining the Pareto front.

5.1. Cost-based formulations

Based on the three models presented in Section 3, we introduce compact MILP formulations for the LoRP that will better allow us to proceed with the analysis. Note that (2-IF) and (MLF) do not ensure that the vehicles return to the depot where they start their routes. Consequently, additional constraints linking the routing variables to the associated depot are required. The f-variables considered in model (3-IF) will now be required in the reformulation of (2-IF) and (MLF). For the three new formulations, the following additional variables are also needed: ι_{ij} represents the remaining length of the route when traversing the edge (i,j) $\forall i,j \in V, i \neq j$. The total travel time of each route is represented by the variables ι_{ij} $\forall i \in D$, $j \in C$, which contains the sum of the travel times of all edges in the respective route.

Reformulated (2-IF). An MILP formulation for the LoRP can be derived from (2-IF) by replacing the objective function (1) with

$$\text{minimize} \quad F_R \sum_{i\in V}\sum_{j\in V\setminus\{i\}} t_{ij}r_{ij} + F_A \sum_{i\in C}\sum_{j\in D} t_{ij}s_{ij} + \sum_{i\in D}\rho_i z_i, \tag{59}$$

removing constraints (6) and (7), and adding (23), (24), (36) plus the following constraints:

$$\text{subject to} \quad f_{ij} \geq r_{ji} \qquad \forall i \in C,\ j \in D, \tag{60}$$

$$f_{ij} \geq r_{ij} \qquad \forall i \in C,\ j \in D, \tag{61}$$

$$r_{jk} \leq 1 - f_{ji} + f_{ki} \qquad \forall i \in D,\ j,k \in C,\ j \neq k, \tag{62}$$

$$r_{jk} \leq 1 + f_{ji} - f_{ki} \qquad \forall i \in D,\ j,k \in C,\ j \neq k, \tag{63}$$

$$\sum_{j\in V\setminus\{i\}} \iota_{ji} - \sum_{j\in V\setminus\{i\}} \iota_{ij} = \sum_{j\in V\setminus\{i\}} t_{ji}r_{ji} + \sum_{j\in D} t_{ij}r_{ij} \qquad \forall i \in C, \tag{64}$$

$$\iota_{ij} \leq T\, r_{ij} \qquad \forall i,j \in V,\ i \neq j, \tag{65}$$

$$\sum_{i\in C}(f_{ij} + s_{ij})d_i \leq \kappa_j z_j \qquad \forall j \in D, \tag{66}$$

$$\sum_{i\in D} z_i \geq \left\lceil \sum_{j\in C} d_j \Big/ \sum_{i\in D} \kappa_i \right\rceil, \tag{67}$$

$$\iota_{ij} \geq 0 \qquad \forall i,j \in V,\ i \neq j. \tag{68}$$

The objective function (59) represents the total cost. The first two terms correspond to the transportation cost, while the third represents the setup cost for the depots. Constraints (60) and (61) ensure that the edges between depots and customers exist only if such customers are allocated to the respective depots. Constraints (62) and (63) ensure the consistency of the defined routes, i.e., they allow tracking the depot each customer is allocated to. Equalities (64) compute the total travel time of the routes, while Inequalities (65) ensure that the maximum length (total travel time) T cannot be exceeded. The capacity constraints for the depots are guaranteed by (66). Valid inequalities for the minimum number of depots required by the system are stated in Constraints (67). Finally, Constraints (68) define the domain of the new decision variables. Note that variables ι_{ij} can be set to 0 or removed from the model when $i,j \in D$, and when $i \in C$, $j \in D$.

Reformulated (3-IF). An MILP formulation for the LoRP can be derived from (3-IF) by replacing the objective function (1) with

$$\text{minimize} \quad F_R \sum_{i\in V}\sum_{j\in V\setminus\{i\}}\sum_{k\in D} t_{ij}r'^{k}_{ij} + F_A \sum_{i\in C}\sum_{j\in D} t_{ij}s_{ij} + \sum_{i\in D}\rho_i z_i, \tag{69}$$

and adding (66)–(68) plus the following constraints:

$$\sum_{j\in V\setminus\{i\}} \iota_{ji} - \sum_{j\in V\setminus\{i\}} \iota_{ij} = \sum_{j\in V\setminus\{i\}}\sum_{k\in D} t_{ji}r'^{k}_{ji} + \sum_{j\in D}\sum_{k\in D} t_{ij}r'^{k}_{ij} \qquad \forall i \in C, \tag{70}$$

$$\iota_{ij} \leq T \sum_{k\in D} r'^{k}_{ij} \qquad \forall i,j \in V,\ i \neq j. \tag{71}$$

The objective function (69) accounts for the total cost of the system in the variable space of (3-IF). Equalities (70) compute the total length (total travel time in our case) of the routes, and Inequalities (71) ensure the maximum length for the routes. As before, variables ι_{ij} can be set to 0 or removed from the model when $i,j \in D$, and when $i \in C$, $j \in D$.

Reformulated (MLF).. To introduce an MILP formulation for the LoRP derived from (MLF), it is necessary to consider an additional set of binary variables. In particular, τ_{ij} is equal to 1 if and only if the edge going from customer $i \in C$ to depot $j \in D$ is in the solution. Using these variables and several sets of additional constraints, it is possible to reformulate the acyclic formulation presented in Nucamendi-Guillén et al. (2022) to include closed routes. The new model can be obtained by replacing the objective function (37) with

$$\text{minimize} \quad F_R\left(\sum_{i\in V}\sum_{j\in C\setminus\{i\}}\sum_{\ell\in L\setminus\{M\}} t_{ij}y_{ij\ell} + \sum_{i\in C}\sum_{j\in D} t_{ij}\tau_{ij}\right) + F_A\sum_{i\in C}\sum_{j\in D} t_{ij}s_{ij} + \sum_{i\in D}\rho_i z_i, \tag{72}$$

and adding (24), (66)–(68) plus the following constraints:

$$\sum_{\ell\in L\setminus\{M\}} x_{i\ell} = \sum_{j\in D} f_{ij} \qquad \forall i \in C, \tag{73}$$

$$\sum_{j\in C}\sum_{\ell\in L\setminus\{M\}} y_{ij\ell} = \sum_{j\in C} \tau_{ji} \qquad \forall i \in D, \tag{74}$$

$$y_{ij\ell} \le f_{ji} \qquad \forall i \in D,\ j \in C, \ell \in L\setminus\{M\}, \tag{75}$$

$$1 + \tau_{ij} \ge f_{ij} + x_{i1} \qquad \forall i \in C, j \in D, \tag{76}$$

$$\sum_{\ell\in L\setminus\{M\}} y_{jk\ell} \le 1 - f_{ji} + f_{ki} \qquad \forall i \in D,\ j,k \in C,\ j \ne k, \tag{77}$$

$$\sum_{\ell\in L\setminus\{M\}} y_{jk\ell} \le 1 + f_{ji} - f_{ki} \qquad \forall i \in D,\ j,k \in C,\ j \ne k, \tag{78}$$

$$\sum_{j\in V\setminus\{i\}} \iota_{ji} - \sum_{j\in V\setminus\{i\}} \iota_{ij} = \sum_{j\in V\setminus\{i\}}\sum_{\ell\in L\setminus\{M\}} t_{ji}y_{ji\ell} + \sum_{j\in D} t_{ij}\tau_{ij} \qquad \forall i \in C, \tag{79}$$

$$\iota_{ij} \le T \sum_{\ell\in L\setminus\{M\}} y_{ij\ell} \qquad \forall i \in V, j \in C,\ i \ne j, \tag{80}$$

$$\iota_{ij} \le T\,\tau_{ij} \qquad \forall i \in C, j \in D, \tag{81}$$

$$\tau_{ij} \in \{0,1\} \qquad \forall i \in C,\ j \in D. \tag{82}$$

The objective function (72) accounts for the total cost of the system in the variable space of (MLF). Equalities (73) ensure that customers appear in the multi-layer network only if they are going to be visited by a vehicle. Constraints (74) establish that for each depot, the number of vehicles departing is equal to the number of vehicles returning. Constraints (75) indicate that the edges between depots and customers may exist only if the customers are to be visited by some vehicle. Constraints (76) ensure that the edge connecting customer i and depot j exists only if customer i is the last in the respective route, and if customer i is allocated to a vehicle departed from depot j. Constraints (77) and (78) ensure the consistency of the defined routes, i.e., they allow tracking the depot each customer is allocated to. Equalities (79) compute the length (total travel time in our case) of the routes, and Inequalities (80) and (81) guarantee the maximum total travel time for the routes. Finally, Constraints (82) state the domain of the new variables. Again, we can set equal to 0, or remove, the variables ι_{ij} $\forall i, j \in D$, and when $i \in C$, $j \in D$. In such a case, Constraints (81) can also be neglected.

The above MILP models presented for the LoRP can be looked at as a contribution in itself. The fact is that to the best of the authors' knowledge, there is no compact model in the literature for this problem. We recall that a set covering formulation solved within a branch-and-price scheme is presented in Arslan (2021). Even though the presented models require larger computing times than the current state-of-the-art algorithms—namely branch-and-price (Arslan 2021) and ALNS (Haghi et al. 2023)—they can be seen as a useful and easy-to-implement tool for researchers and practitioners that face similar problems with small and medium-sized instances. Additionally, as will be presented in Section 6.2, the proposed models can provide new upper bounds and solve to proven optimality several instances with up to 50 customers. The proposed formulations can be easily extended to other objective functions to model realistic problems. Moreover, it is helpful for the latency-cost tradeoff analysis we wish to conduct.

5.2. Finding the Pareto front

If a latency objective function is added to the above models, we obtain bi-objective MILPs. There are several popular tools for finding Pareto solutions in such a case. We adopt the ε-constrained method, which considers one objective function and sets the other as a constraint. By performing an adequate parametric analysis on the bound set for that additional constraint, it is possible to obtain the Pareto front or part of it.

In the context of our problem, for any of the proposed models, we proceed as follows:

1. The above-introduced single-objective model is solved to optimality considering the cost-based objective function. This leads to an extreme ray of objective space. By minimizing the latency while ensuring that the optimal cost does not increase, we obtain one point of the so-called payoff table, which is in fact one extreme of the Pareto front. We denote it by A = $(cost^*, lat_A)$.
2. The above model with latency as the single objective function is solved to optimality, which leads to another extreme ray in the objective space. By optimizing the cost while ensuring that the latency does not increase with respect to the optimal value, we obtain the other point in the payoff table, i.e., the other extreme of the Pareto front. We denote it by B = $(cost_B, lat^*)$.

3. The single-objective model considering the latency-based objective function is then iteratively solved to optimality for a certain number, say `iter`, of iterations. In each iteration constraint (83) is added, where ε represents the maximum value of cost imposed. At each iteration ε is equal to its value at the previous iteration minus $\Delta cost$, with $\Delta cost = \frac{cost_B - cost^*}{\texttt{iter}}$. In the first iteration ε is set equal to $cost_B - \Delta cost$.

$$F_R \sum_{i \in V} \sum_{j \in V \setminus \{i\}} t_{ij} r_{ij} + F_A \sum_{i \in C} \sum_{j \in D} t_{ij} s_{ij} + \sum_{i \in D} \rho_i z_i \leq \varepsilon \tag{83}$$

Notice that Constraint (83) uses the expression of total cost under the variable space of (2-IF); nevertheless, it can be trivially obtained for any of the proposed formulations from expressions (69) and (72), for (3-IF) and (MLF), respectively.

6. Computational experiments

The MILP models discussed in the previous sections were implemented in C++, and all the experiments were executed on an AMD EPYC 7262 8-Core Processor with 32 GB RAM and 16 physical cores, under Rocky Linux 8.10 (Green Obsidian) operating system. Gurobi 11.0.0, under its default parameter configuration and with a single thread setting, was used as the solver. For the benchmark analysis presented in Section 6.1 a time limit of 7200 seconds was considered for each instance of the model. The tests conducted for the LoRP (see Section 6.2) considered a time limit of 10,800 seconds, as done previously by Haghi et al. (2023) for the same problem. While for all the other experiments no time limit was considered.

The source code of the methods presented in this paper will be publicly available upon request to the corresponding author.

The set of instances used for executing the experiments related to the LLRP corresponds to all the instances with up to 100 customers introduced by Moshref-Javadi and Lee (2016) that were based on the instances introduced by Barreto (2004) and Prins et al. (2006) for the LRP. In total, 32 base instances were considered. Nevertheless, the instances introduced by Moshref-Javadi and Lee (2016) do not contain values for F_A, F_R, and R. Accordingly, we followed the procedure by Arslan (2021) for generating these parameters, although we consider a more diverse set. The following values were considered for each parameter: $R \in \{5, 10, 15, 20, 30, 40, 50\}$, $F_A \in \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0\}$, and $F_R = 1$. Thus, 77 combinations of parameters were obtained for each base instance, leading to a total of 2464 instances. The reason for considering $F_R = 1$ is that by keeping this factor fixed and varying the allocation factor, we can better understand the impact of changing the relation between allocation and routing times. Note that for the managerial insights provided in Section 6.3, additional changes in some parameters are considered, namely involving the potential number of depots to select, p, and the capacity of the vehicles, Q.

For the experiments associated with the LoRP (Section 6.2) we consider the same instances for which the authors in Haghi et al. (2023) present detailed results. This corresponds to ten base instances with up to 50 customers from the so-called Prodhon data set (Prins et al. 2006), and 12 base instances with up to 40 customers from the Akca data set (Akca et al., 2009). The parameters are: $R \in \{0, 10, 20, 30, 40, 50\}$, $F_A = 0$, $F_R = 1$, and $T \in \{100, 125, 150\}$. Notice that some instances are infeasible under the mentioned parameter configurations. A total of 354 feasible instances were found and thus analyzed. Additionally, as done by Haghi et al. (2023), the depots were considered uncapacitated and there is no limit on the number of depots to open and vehicles to use.

All the instances in this paper consider parameter t_{ij} equal to the respective Euclidean distance, although the experiments presented in Section 6.2 consider a scaling factor as done by Arslan (2021) and Haghi et al. (2023).

6.1. Performance of the formulations and lower bounds

In this section, we report on the results of a series of tests conducted using the MILP models introduced in Section 3 and the lower bounds discussed in Section 4. Also, regarding the lower bounds, we recall that in addition to LB1 and LB2 it is possible to consider the linear programming (LP) relaxation of the proposed formulations. We denote them by LP-2IF, LP-3IF, and LP-MLF, respectively. Given the large number of instances, the detailed results obtained for each of them can be accessed by the readers in the Electronic Supplementary Materials (ESM).

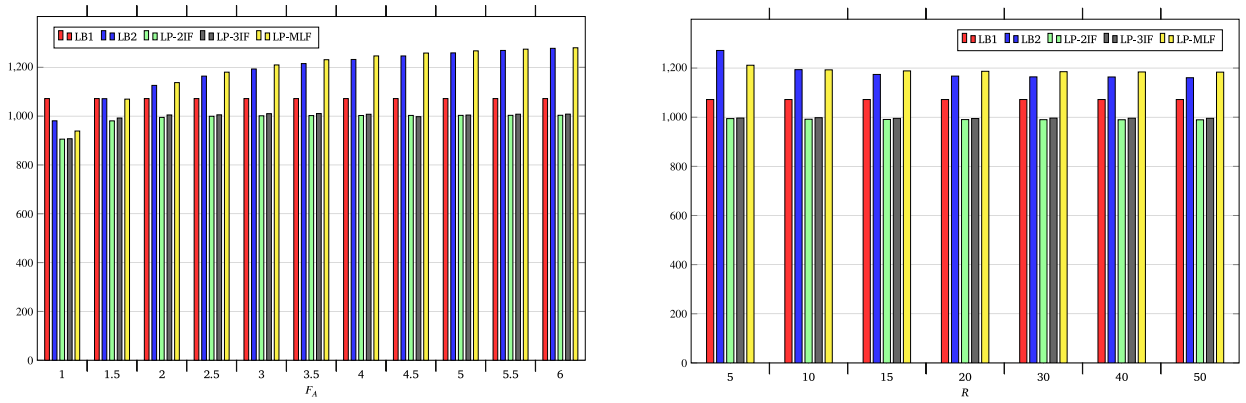
Table 2 summarizes the results obtained for each base instance. Results are aggregated for the 77 combinations of F_A and R. The first column identifies the base instance. The first value in the name denotes the number of customers, while the second indicates the number of depots. For each base instance, the following three columns are also reported: $|K|$ is the number of available vehicles, $\#opt$ regards the number of instances solved to proven optimality within the time limit by at least one of the three formulations; $\#feas$ concerns the number of instances for which the solver can find a feasible solution (not optimal) within the time limit with at least one of the three formulations; $\#noSol$ is the number of instances for which the solver is not able to find a feasible solution within the time limit with any of the three formulations. Additionally, for each formulation, five columns are presented: $time$ refers to the average computing time in seconds (this column is also reported for each lower bound); $mipgap$ concerns the average percentage gap reported by the solver upon termination for the instances not solved to proven optimality; and the already defined columns $\#opt$, $\#feas$, and $\#noSol$. Finally, for each instance and each lower bound the following additional columns are reported: gap_B contains the average percentage gap of the value obtained by the corresponding lower bound ($Bound$) w.r.t. the best feasible solution value obtained after running the three formulations until the time limit ($Best$), computed as $gap_B = 100(Best - Bound)/Best$ (see the ESM for all information about $Best$ and $Bound$); $\#tight$ is the number of instances for which the corresponding lower bound is the tightest one. Finally, the last row in Table 2 (Global results) presents the global average values for all the mentioned columns but for $\#opt$, $\#feas$, $\#noSol$, and $\#tight$, where it reports the sum of the respective values. Note that the values of $mipgap$ associated with each formulation might be associated with a different number of instances. Consequently, these values are not comparable between formulations.

Table 2
Aggregated results for each base instance: formulations and lower bounds.

		(2-IF)								(3-IF)										(MLF)										(LB1)			(LB2)			(LP-2IF)			(LP-3IF)			(LP-MLF)		
Instance	$\|K\|$	#opt	#feas	#noSol	time	mipgap	#opt	#feas	#noSol	time	mipgap	#opt	#feas	#noSol	time	mipgap	#opt	#feas	#noSol	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight	time	gap_B	#tight
20-5-1	5	77	—	—	8.3	—	77	—	0	39.9	—	77	—	0	1.8	—	77	—	0	<0.1	23.7	5	0.1	9.9	7	0.1	15.1	0	0.5	14.4	0	<0.1	9.2	65
20-5-1b	3	77	—	—	59.9	—	77	—	0	237.8	—	77	—	0	2.9	—	77	—	0	<0.1	31.2	4	0.3	8.7	17	0.1	23.5	0	0.5	23.4	0	0.1	9.0	56
20-5-2	5	77	—	—	2.9	—	77	—	0	14.4	—	77	—	0	0.8	—	77	—	0	<0.1	22.5	7	0.2	8.1	28	0.1	16.6	0	0.5	15.6	7	<0.1	8.0	35
20-5-2b	3	77	—	—	93.3	—	77	—	0	324.4	—	77	—	0	2.9	—	77	—	0	<0.1	35.9	5	0.4	13.6	53	0.1	31.4	0	0.7	31.0	0	0.1	16.8	19
Gaskell-21-5	4	77	—	—	73.9	—	77	—	0	336.9	—	77	—	0	0.7	—	77	—	0	<0.1	36.2	3	0.2	9.9	22	0.1	24.5	0	0.8	23.8	0	0.1	10.4	52
Min-27-5	4	77	—	—	60.5	—	77	—	0	283.6	—	77	—	0	14.1	—	77	—	0	<0.1	18.6	0	1.6	3.5	77	0.2	22.4	0	2.5	22.1	0	0.2	11.7	0
Gaskell-29-5	4	77	—	—	4825.1	6.3	30	47	0	5699.9	8.6	20	57	0	36.4	—	77	—	0	<0.1	29.7	6	1.1	14.6	30	0.2	33.6	0	2.6	33.0	0	0.3	16.4	41
Gaskell-32-5b	3	77	—	—	6041.1	13.2	16	61	0	6447.6	17.1	9	68	0	22.2	—	77	—	0	<0.1	40.8	3	6.5	4.8	45	0.3	32.1	0	4.5	31.9	0	0.7	9.6	29
Gaskell-36-5	4	77	—	—	6214.9	20.8	11	66	0	6588.3	27.0	8	69	0	4.0	—	77	—	0	<0.1	48.4	3	1.1	12.1	22	0.4	37.4	0	6.2	37.4	0	0.3	12.9	52
50-5-1	12	77	—	—	2165.2	0.7	75	2	0	5030.3	1.6	27	50	0	23.9	—	77	—	0	<0.1	10.5	5	1.7	3.5	4	1.0	7.0	0	13.5	6.5	0	0.5	3.5	68
50-5-1b	6	77	—	—	6809.5	8.4	5	72	0	6857.9	11.8	4	73	0	374.4	—	77	—	0	<0.1	21.8	18	6.2	18.1	54	1.1	33.5	0	21.4	32.8	0	0.9	20.3	5
50-5-2	12	77	—	—	5058.5	1.2	32	45	0	5771.8	2.6	18	59	0	167.9	—	77	—	0	<0.1	10.5	14	2.7	10.1	0	1.3	13.8	0	15.1	13.1	0	0.6	9.7	63
50-5-2b	6	77	—	—	6291.1	12.5	10	67	0	6462.0	20.6	10	67	0	99.0	—	77	—	0	<0.1	29.7	5	7.4	7.5	61	1.4	27.1	0	18.0	27.0	0	0.9	10.3	11
50-5-2BIS	12	77	—	—	1848.1	0.2	76	1	0	6023.0	0.4	14	63	0	291.5	—	77	—	0	<0.1	3.1	3	6.7	0.9	77	0.8	1.7	3	11.9	1.7	3	0.7	0.9	3
50-5-2bBIS	6	77	—	—	6077.7	3.7	13	64	0	6132.5	6.4	12	65	0	13.0	—	77	—	0	<0.1	15.3	3	13.3	0.3	77	1.1	10.0	3	17.7	10.0	3	1.6	1.2	3
50-5-3	12	77	—	—	5438.3	2.7	21	56	0	5763.5	3.7	19	58	0	1422.8	—	77	—	0	<0.1	9.6	77	1.9	20.8	0	1.0	24.8	0	15.4	23.6	0	0.6	17.8	0
50-5-3b	6	77	—	—	6310.9	10.6	10	67	0	6448.2	16.6	9	68	0	243.7	—	77	—	0	<0.1	27.1	12	17.2	12.7	41	1.3	35.0	0	22.4	34.8	0	0.9	14.7	24
Christ-50-5	6	77	—	—	6643.9	13.3	6	71	0	6669.8	23.8	6	71	0	239.1	—	77	—	0	<0.1	33.5	10	10.9	20.8	33	1.4	43.9	0	20.0	43.6	0	1.1	20.1	34
Christ-75-10	9	62	15	—	6761.5	22.7	6	71	0	7128.0	32.9	3	61	13	2939.4	0.4	62	15	0	<0.1	25.5	5	23.2	14.3	5	5.1	28.3	0	879.8	28.0	0	11.6	8.9	67
100-5-1	24	6	66	5	6738.6	4.0	5	67	5	6786.1	9.8	5	30	42	6675.4	0.9	6	44	27	<0.1	5.1	14	159.6	6.1	0	12.9	6.6	0	374.5	6.1	0	10.3	4.9	63
100-5-1b	12	23	54	—	6833.6	9.4	4	73	0	6897.6	14.4	4	66	7	5461.4	1.5	23	38	16	<0.1	14.4	5	288.4	5.7	12	12.5	13.2	0	400.5	13.0	0	25.3	5.5	60
100-5-2	24	6	71	—	6645.6	8.5	6	70	1	6718.4	18.3	6	59	12	6659.1	1.9	6	66	5	<0.1	8.0	77	15.4	16.5	0	11.2	17.7	0	345.5	15.7	0	4.7	13.2	0
100-5-2b	11	29	48	—	6741.0	33.9	5	72	0	6796.9	33.4	5	59	13	5050.4	1.8	29	37	11	<0.1	28.8	6	99.0	15.0	9	16.6	31.6	0	545.5	31.2	0	17.6	14.7	62
100-5-3	24	6	68	3	6666.8	9.4	6	67	4	6811.5	16.5	5	62	10	6693.9	3.5	6	56	15	<0.1	8.4	77	32.8	35.0	0	10.6	37.1	0	328.0	35.0	0	4.9	33.3	0
100-5-3b	11	7	70	—	6747.8	43.1	5	72	0	6877.6	42.7	4	61	12	6883.9	19.2	6	50	21	<0.1	27.1	38	292.3	31.8	39	16.1	46.4	0	662.5	46.0	0	11.9	32.5	0
100-10-1	26	6	71	—	6764.1	10.6	6	71	0	7200.0	39.5	0	51	26	6670.6	4.6	6	65	6	0.1	7.7	77	20.5	33.4	0	9.0	34.1	0	2489.2	31.4	0	3.6	26.2	0
100-10-1b	12	5	72	—	6863.4	38.8	4	71	2	7200.0	52.4	0	54	23	7013.9	10.7	4	58	15	0.1	28.0	20	63.6	29.4	23	18.8	44.9	0	4136.0	44.2	0	13.2	26.2	34
100-10-2	24	14	62	1	6654.2	9.0	6	70	1	7148.2	21.7	3	53	21	6051.5	2.4	14	50	13	<0.1	11.1	14	27.5	13.2	0	10.2	15.5	0	3376.5	14.6	0	4.8	10.4	63
100-10-2b	11	18	59	—	6746.4	29.6	5	72	0	7200.0	38.4	0	55	22	6011.5	3.0	18	51	8	<0.1	30.8	6	299.6	12.1	47	17.8	30.1	0	4579.7	30.0	0	14.4	12.7	24
100-10-3	24	12	65	—	6671.7	6.1	6	71	0	7187.5	29.8	2	55	20	6331.1	3.0	12	51	14	<0.1	8.6	77	47.4	22.8	0	8.3	23.5	0	2509.0	20.8	0	4.4	18.5	0
100-10-3b	11	6	71	—	6765.2	37.0	5	72	0	7200.0	43.2	0	65	12	7088.5	7.0	3	63	11	<0.1	29.3	6	102.8	23.9	2	14.8	37.1	0	3640.6	36.3	0	13.8	20.6	69
Christ-100-10	8	21	56	—	6948.4	55.7	3	74	0	7200.0	56.8	0	63	14	5666.5	6.9	21	54	2	<0.1	40.9	5	323.1	29.4	8	18.4	50.4	0	4619.0	49.6	0	49.3	19.4	64
Global results (2464 inst.)	1607	848	9	4955.4	17.9	839	**1612**	**13**	5421.4	22.7	655	1562	247	**2754.9**	**5.2**	**1602**	698	164	**<0.1**	22.6	610	58.6	14.6	793	6.1	26.6	6	908.1	25.7	13	6.2	**14.0**	**1066**

Table 3
Comparison of model performance by instance size.

		(MLF)		(2-IF)		(3-IF)	
Instance size	#	#opt	time	#opt	time	#opt	time
Small ($\|C\| < 50$)	693	**693**	**9.5**	519	1931.1	499	2219.2
Medium ($50 \leq \|C\| < 100$)	770	**755**	**581.5**	254	5340.5	122	6228.7
Large ($\|C\| = 100$)	1001	**154**	**6327.5**	66	6752.8	34	7017.2
All	2464	**1602**	**2754.9**	839	4955.4	655	5421.4



(a) (b)

Fig. 3. Value of the lower bounds for different values of (**a**) F_A, (**b**) R.

Nevertheless, *mipgap* provides valuable information about the quality of the upper bounds obtained by each model. The same holds for the Global result associated with the mentioned column, which is computed without considering the instances for which each formulation is not able to find a feasible solution. For the compact formulations, we additionally provide aggregated results by instance size in Table 3, where # is the number of instances.

Lower bounds. According to the results, LP-MLF provides the tightest lower bound for most of the instances (1066 instances), followed by LB2 (793 instances), LB1 (610 instances), LP-3IF (13 instances), and LP-2IF (6 instances). Note that in six out of 2464 instances the five lower bounds are equal; this situation occurs in two particular base instances when the value of R is relatively large and $F_A = 1$. A more detailed analysis (see the ESM) reveals that the average deviation w.r.t. the tightest lower bound value obtained for each instance is 4.7%, 5.4%, 14.2%, 18.1%, and 19% for LP-MLF, LB2, LB1, LP-3IF, and LP-2IF, respectively (values and presented in increasing order). All the proposed lower bounds but LP-3IF require relatively small computing times (less than a minute on average). In the case of LP-3IF the average computing time is 908 seconds, which is not compensated by the quality of the lower bound obtained. Additionally, for 26 instances, the solver is not able to solve LP-3IF within the time limit.

The quality of the lower bounds for different values of F_A and R is presented in Fig. 3. Each bar stands for the average value of the lower bound for the tested instances. According to the results, LB1 does not seem sensitive to the values of R and F_A, since in our experiments it always holds that $F_A \geq F_R$ and thus, the single-customer routes will always be as efficient or more than the direct allocation. LB1 is tight (relative to the other lower bounds) for small values of F_A, i.e, when direct allocation is more likely to happen. LB2 and LP-MLF behave similarly for different values of F_A and R. Both lower bounds are particularly tight (w.r.t. LB1, LP-2IF, and LP-3IF) when direct allocation is not likely to happen (large values of F_A and small values of R). A visible difference in favor of LB2 is found when $R = 5$.

Optimization models. It is worth noticing that the LP relaxation of (MLF) is often stronger than that of (2-IF) and (3-IF). This superiority was observed in 2436 and 2381 instances for (2-IF) and (3-IF), respectively (see the ESM). The average deviation of the optimal value of LP-MLF w.r.t. the optimal value of LP-2IF and LP-3IF is 19.5% and 18.2%, respectively.

Looking into Tables 2 and 3, we see that (MLF) provides 1602 optimal solution values, followed by (2-IF) and (3-IF) with 839 and 655 instances solved to proven optimality, respectively. Regarding the number of upper bounds, it is possible to state that (2-IF) is the dominant formulation, failing to provide an upper bound for just 13 instances, while (MLF) and (3-IF) fail to provide a feasible solution for 164 and 247 instances, respectively. There are 9 instances with $|C| = 100$ for which none of the proposed formulations is able to find a feasible solution within the time limit (see the ESM). Considering the upper bounds not proved to be optimal, the average mipgap upon termination reported by the solver using each formulation is the following: 5.2% for 698 instances using (MLF), 17.9% for 1612 instances using (2-IF), and 22.7% for 1562 instances using (3-IF).

Regarding the computing times, (MLF), (2-IF), and (3-IF) require, on average, 2755, 4955, and 5421 seconds to solve the considered instances, respectively. Thus, on average, (2-IF) and (3-IF) are almost two times slower than (MLF). In a more detailed analysis (ESM) comparing the computing times associated with each instance, we found that for 1503 instances, (MLF) is the fastest among the three formulations. Additionally, for 104 instances, (2-IF) requires less computing time than the other two formulations; for 857 instances, the three formulations reach the time limit.

Another interesting analysis emerges when the instances are grouped by the number of customers ($|C|$). We consider three cases: *i*) small ($|C| < 50$), *ii*) medium ($|C| \in [50, 100[$), and *iii*) large ($|C| = 100$). Observing Table 3 we conclude that using (MLF) we could solve to proven optimality all the 693 small-sized instances, and almost all the medium-sized instances (755 out of 770) within the time limit. On the other hand, using (2-IF) (resp. (3-IF)) we could solve to proven optimality 519 (resp. 499) and 254 (resp. 122) small and medium instances, respectively, while for 174 (resp. 194) and 516 (resp. 648) small and medium instances, respectively, it is not able to prove the optimality of the solutions within the time limit. Regarding the computing times, there is also a remarkable difference between the formulations; (MLF) requires on average 10 and 582 seconds for solving small- and medium-sized instances, respectively, while (2-IF) (resp. (3-IF)) requires on average 1931 (resp. 2219) and 5341 (resp. 6229) seconds. When large instances are considered, in most cases, the three formulations reach the time limit. Nevertheless, the number of instances solved to proven optimality by (MLF) is about 2 and 3 times larger than that for (2-IF) and (3-IF) (154 against 66 and 34 instances, respectively).

Looking into the solutions obtained in more detail (ESM) we observe that the upper bound value (Lat) obtained by (MLF), (2-IF), and (3-IF) matches $Best$ for 2244, 1428, and 1084 instances, respectively. The average deviation of Lat w.r.t. $Best$, computed for each instance as $100(Lat - Best)/Best$, obtained by (MLF), (2-IF), and (3-IF) is 0.1%, 4.5%, and 6.9%, respectively. When comparing the results obtained by each formulation it is possible to observe the following.

- (MLF) versus (2-IF). For 1017 (MLF) finds a better solution value than (2-IF), for 1235 instances both formulations obtain the same solution value, for 10 instances neither of the formulations finds an upper bound, while (2-IF) finds a better solution value than (MLF) for 202 instances.
- (MLF) versus (3-IF). We observe that for 1223 (MLF) finds a better solution value than (3-IF), for 1066 instances both formulations obtain the same solution value, for 118 instances neither of the formulations finds an upper bound, and for 57 instances (3-IF) finds a better solution value than (MLF).
- (2-IF) versus (3-IF). We found that for 1124 (2-IF) finds a better solution value than (3-IF), for 1051 instances both formulations obtain the same solution value, for 10 instances none of the formulations finds a feasible solution, and for 279 instances (3-IF) finds a better solution value than (2-IF).

In summary, we can state that (MLF) dominates—in terms of solution quality and computing time—both (2-IF) and (3-IF), and that (2-IF) dominates (3-IF).

As a final remark, we would like to mention that using multiple threads the best formulation, namely (MLF), is able to solve 1776 instances to proven optimality (versus 1602 using a single thread) with an average computing time equal to 2323.8 seconds (versus 2754.9 seconds when using a single thread). Using multiple threads, (MLF) can solve to proven optimality all the instances with up to 75 customers within the time limit. We present the results using a single thread to provide a fair comparison point for future researchers. Nevertheless, we note that in the managerial insights section (Section 6.3) an additional computational effort is considered to explore as much as possible the presented formulations and thus obtain as many optimal solutions as possible.

From a managerial perspective, the results indicate that provably optimal solutions can be obtained within reasonable computing times using (MLF) for instances with up to 75 customers. Such sizes are representative of many humanitarian and urban logistics applications, making the proposed model a practical decision-support tool that can be solved with off-the-shelf optimization solvers. At the same time, the computational effort required for larger instances motivates the development of more advanced exact and heuristic solution methods.

6.2. Performance on the standard LoRP

In this section, the results obtained by the three compact formulations are compared against those obtained by the branch-and-price (BP) algorithm of Arslan (2021), the ALNS heuristic by Haghi et al. (2023) and the MVNS heuristic by Zhou et al. (2026). It is to be noted that the authors in Zhou et al. (2026) do not perform experiments using the Akca data set. According to the single-thread score reported in Passmark [2], the computers used in Haghi et al. (2023) and Zhou et al. (2026) are 1.12 and 1.13 times faster than ours, respectively. Thus, the computing times reported in Haghi et al. (2023) for BP and ALNS, and in Zhou et al. (2026) for MVNS, are scaled accordingly.

Table 4 reports the aggregated results of the following metrics. $\#\text{opt}_{MILP}$ is the number of instances solved to proven optimality by at least one of the compact formulations introduced in Section 5. $\#\text{BKS}_{+}$ is the number of new best-known solutions (BKSs) found by performing these experiments. #UB is the number of instances for which an upper bound is found (this column is not reported for ALNS and MVNS since they find an upper bound for each instance for which they are executed). #BKS is the number of instances for which the upper bound obtained matches the BKS. Column dev represents the percentage deviation of the upper bound to the BKS. Like in the previous section, the detailed results for each instance can be accessed in the supplementary materials.

[2] https://www.cpubenchmark.net/

Table 4
Aggregated results for the standard LoRP.

				BP				ALNS			MVNS			(2-IF)					(3-IF)					(MLF)				
Data set	#	$\#opt_{MILP}$	$\#BKS_{+}$	#UB	#BKS	dev	time	#BKS	dev	time	#BKS	dev	time	#opt	#UB	#BKS	dev	time	#opt	#UB	#BKS	dev	time	#opt	#UB	#BKS	dev	time
Akca	192	107	17	**192**	**175**	**<0.1**	344.4	115	0.5	**35.2**	—	—	—	**106**	**192**	174	0.1	5339.8	92	191	170	0.1	6053.6	64	187	148	0.5	7553.2
Prodhon	162	14	72	160	37	<0.1	2518.6	22	<0.1	81.7	63	**<0.1**	**39.4**	**14**	144	**88**	5.5	10046.0	12	138	76	13.3	10123.1	4	111	75	10.2	10611.5
All	354	121	89	352	212	**<0.1**	1339.3	137	0.3	**56.5**	—	—	—	**120**	336	**262**	2.4	7493.5	104	329	246	5.7	7915.9	68	298	223	4.1	8952.7

According to the results, 121 out of the 354 instances were solved to proven optimality by at least one of the proposed formulations within the time limit. Among the proposed models, (2-IF) is the best compact formulation for solving the LoRP. Using this model, it was possible to solve 120 instances to proven optimality within the time limit versus 104 for (3-IF) and 68 for (MLF). Additionally, the use of (2-IF) yields the shortest computing times, and the highest number of upper bounds and best-known solutions among the three formulations. It is interesting to note that (MLF) performed the best when solving the LLoRP but the worst when solving the LoRP.

By looking into the detailed results, we found that for 92 instances (81 from Prodhon and 11 from Akca), the compact formulations obtain an upper bound smaller than the solution value reported as optimal in Haghi et al. (2023). These differences may occur due to implementation issues related to the BP algorithm or to incorrect values being reported for some reason. Although several values classified as optimal by Haghi et al. (2023) are actually sub-optimal, the deviation w.r.t. the values we obtained is negligible (0.02% on average), and does not affect the conclusions drawn by Arslan (2021) for the LoRP. Additionally, for six Akca instances, the solution value reported by BP and ALNS is smaller than the optimal solution value reported by the compact formulations. Again, these cases may occur due to incorrect values being reported for some reason. In synthesis, the formulations we introduced in Section 5 provide a new best upper bound for 89 instances out of the 354 reported in Table 4.

Irrespective of the above, the three algorithms leading to the benchmarking results (Arslan 2021, Haghi et al. 2023 and Zhou et al. 2026) are more efficient than a plain use of the compact formulations we introduce in this paper. This is true in terms of the computing time. Moreover, in general, the solutions reported by those authors are of better quality. Nevertheless, (2-IF) is the method that finds the highest number of BKSs considering the six contenders.

The main conclusion drawn from the previous experiments is that the three newly proposed formulations are able to provide competitive results (in terms of solution quality) in reasonable computing times for the LoRP. Additionally, from a managerial perspective the proposed models constitute a valuable contribution to the existing knowledge because they can be easily extended to other objective functions and constraints, a flexibility that is often required in practical applications.

6.3. Managerial insights

The objective of this section is to provide a comprehensive analysis of the LLoRP and discuss its managerial implications in practical applications. To this end, the section is organized around a series of managerial questions related to key design decisions and model parameters.

First, we demonstrate the relevance of hybrid transportation systems that combine routing and direct allocation, highlighting their advantages over purely routing-based approaches (Section 6.3.1). Next, we analyze how different parameters and assumptions affect both the quality and structure of the optimal solutions, including customer waiting times, depot-location decisions, and the allocation of customers to routing or direct service. In Section 6.3.2 we examine the impact of the coverage range (R) and direct allocation factor (F_A), while in Section 6.3.3 we investigate the effects of the maximum number of depots that can be opened (p) and vehicle-capacity constraints. Finally, in Section 6.3.4 we look into the tradeoff between latency and cost, emphasizing the value of considering latency-oriented objectives in location-or-routing problems. We conclude with a summary of the main findings and their managerial implications in Table 7.

All the experiments conducted in this section, apart from those presented in Section 6.3.4 (the bi-objective framework analysis), were performed using model (MLF), which presented the best performance among the three proposed formulations in Section 3.

We focus the analysis of this section on the base instances for which it is possible to solve to proven optimality all the 77 resultant LLoRP instances —related to the combinations of F_A and R. As presented in Section 6.1, all the instances with up to 50 customers were solved to proven optimality within the time limit. We also considered solving larger instances but without imposing a time limit. Nevertheless, for just three additional base instances, the 77 resulting instances were solved to prove optimality without facing memory issues. Hence, we consider all the base instances with $|C| \leq 50$ plus instances Christ-75-10, 100-10-1b and 100-10-2b. In total 21 base instances are involved, i.e., 1617 instances if we consider the 77 combinations of F_A and R. Unless it is explicitly mentioned, the results reported next refer to those instances. The reader can find all the detailed results mentioned in this section in the ESM.

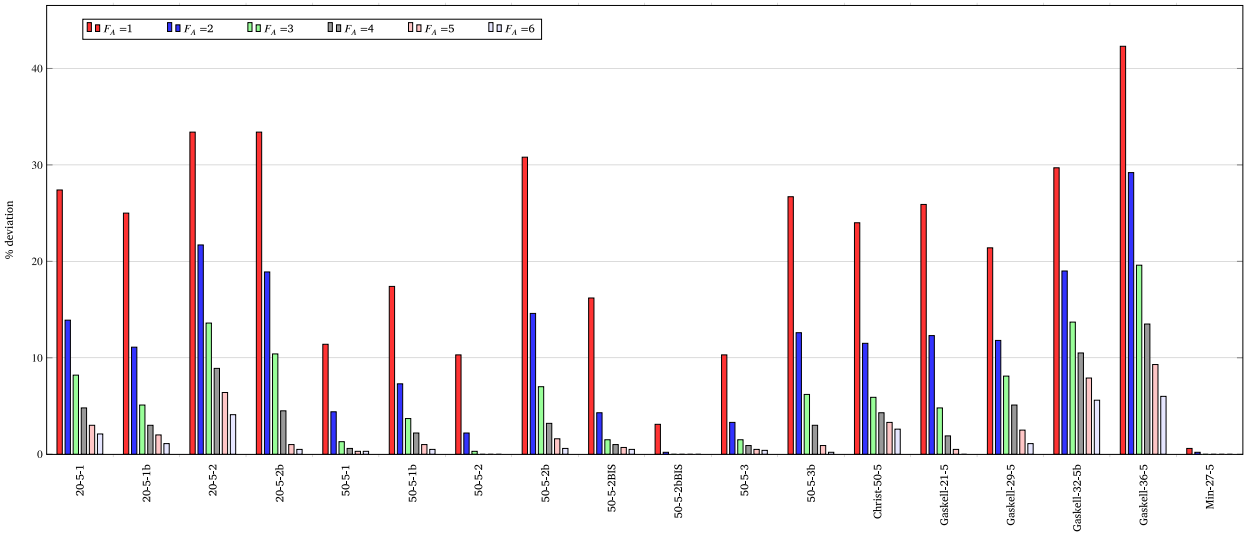
6.3.1. The relevance of integrating direct allocation into latency location-routing problems

In this section, we investigate the value of incorporating direct service into the latency location-routing setting. Specifically, we seek to answer the following managerial questions: *Is it worthwhile to offer direct service in addition to vehicle-based operations? How much can customer waiting times be reduced by doing so? Under which operating conditions does direct service provide significant benefits?*

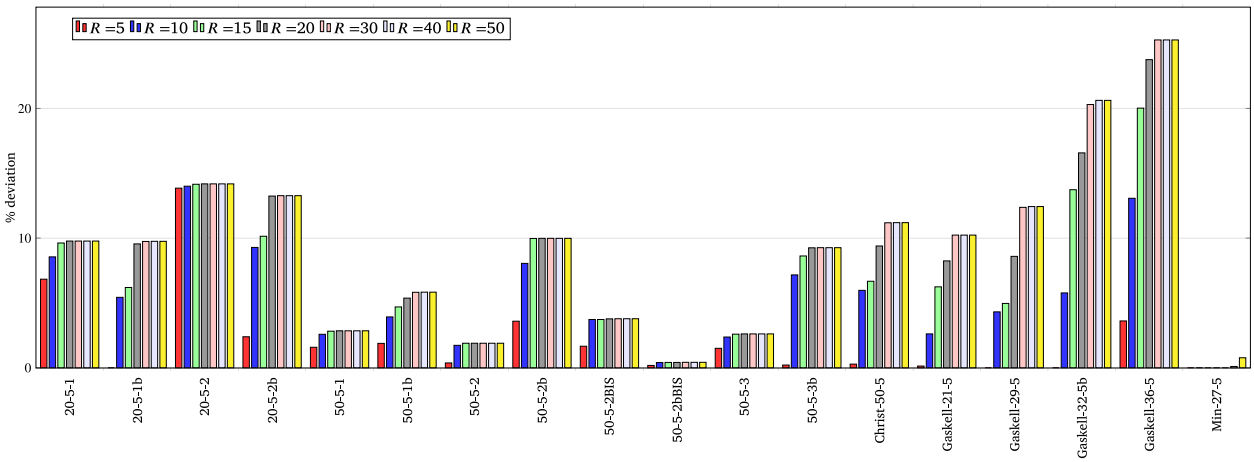
For this analysis, we consider the 18 LLRP base instances with up to 50 customers that have been solved to proven optimality using the formulations presented in Nucamendi-Guillén et al. (2022). Thus, the optimal solution values obtained for each base instance with the 77 combinations of parameters (F_A and R) are compared to the optimal solution value of the LLRP (no direct allocation allowed).

For each instance and each pair (F_A,R) we consider the percentage deviation between the optimal solution value of the LLoRP, say Lat_{LLoRP}, and the optimal solution value of the LLRP, say Lat_{LLRP}: $Dev = 100(Lat_{LLRP} - Lat_{LLoRP})/Lat_{LLRP}$. This is the percentage saving in latency by allowing direct allocation. This value is insightful because, for all the instances, we set $F_R = 1$. Thus, in the worst case (very large values for F_A), the problem reduces to the LLRP (every customer is served by a vehicle as part of a route).

The aggregated results obtained for each base instance are drawn in Fig. 4. The global average saving in latency considering all the instances and the 77 possible combinations of F_A and R is 7.17%. The maximum and minimum average savings in latency are observed for instances Gaskell-36-5 (19.47%) and Min-27-5 (0.12%), respectively. The small savings obtained for Min-27-5 can be



(a) By direct allocation factor F_A.



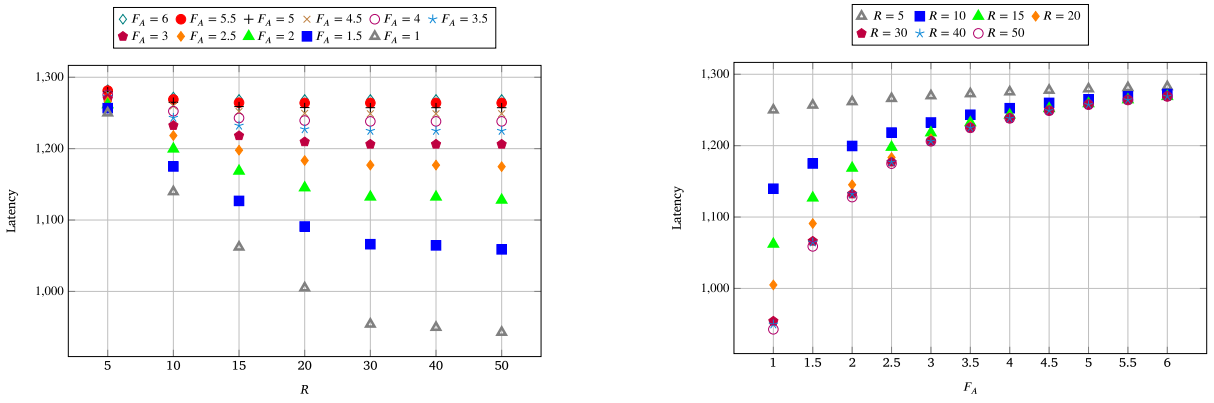
(b) By coverage range R.

Fig. 4. Saving in total customer waiting time (%) achieved by the LLoRP with respect to the corresponding LLRP (no direct service), for each base instance.

explained by the magnitude of the distances in this instance. The fact is that in such a case, the values of R are small compared to the values of the distances and thus, a small percentage of customers is within the coverage range of the depots. For these instances, when we observe Fig. 4(a), we realize that even for $F_A = 1$ the savings are close to 0%. Similarly, in Fig. 4(b), we observe that only for large values of R, savings are reached, although small ones (for $R < 40$ the savings are 0%).

For 11 base instances, the average reduction in latency is over 5%, and just for two instances, it is under 1%. Even for the extreme cases where direct allocation is not likely to happen (say when R is small and F_A is large), it is possible to obtain non-negligible savings. When $R = 5$ (small coverage range) we observe an average deviation equal to 2.12%. On the other hand, when $F_A = 6$ (the direct allocation factor is 6 times larger than the routing factor F_R) the average deviation is equal to 1.41%. None of these values is negligible. Combining both extreme cases the average reduction in latency is equal to 0.62%. In practice, these extreme cases can represent situations where *i*) the customers are not willing (or not able) to walk long distances to receive the service, and *ii*) when moving towards the facility is considerably slower compared to wait for the delivery service (e.g., walking pace versus motor-vehicle speed). Both cases can be illustrated when considering that the customers are children, pregnant women, the elderly, or injured people willing to receive medical services.

In synthesis, the results provide evidence that clear gains in terms of waiting times can be achieved by considering both direct allocation and routing as options for serving the demand nodes. In practical terms, these savings are obtained without deploying any additional resource: the depots and the fleet are exactly those of the corresponding LLRP, and the gain comes only from letting suitably located customers access the facility directly. In a post-disaster distribution operation, a 7% reduction of the total waiting time represents time that is actually subtracted from the deprivation of the affected population; in commercial last-mile delivery, it translates one-to-one into shorter promised lead times at unchanged fleet cost. Moreover, part of the benefit stems from capacity



(a) Latency vs R (curves represent different values of F_A).

(b) Latency vs F_A (curves represent different values of R).

Fig. 5. Impact of the coverage range R and the direct allocation factor F_A on the optimal latency.

release: when nearby customers collect their orders, vehicle capacity is freed for the remaining ones, allowing more efficient routes—an effect that persists even when direct access is slow ($F_A = 6$) or the coverage range is small ($R = 5$).

6.3.2. *The effect of covering range R and direct allocation factor F_A over latency*

In a LoRP setting, increasing the coverage range R is expected to improve system performance, whereas increasing the direct allocation factor F_A should have the opposite effect. In practice, a larger R may result from investments in longer-range drones or from customers being able to travel farther to receive service. Conversely, higher values of F_A represent more challenging access conditions, such as those faced by elderly, injured, or vulnerable individuals, or by customers located in difficult terrains (UNWFP, 2016). This section addresses the following managerial questions: *How do changes in R and F_A affect the structure of the optimal solutions? Is it worthwhile to increase the coverage range? Does direct service remain beneficial under unfavorable access conditions?*

The instances used in experiments reported next correspond to the 77 combinations of R and F_A for each of the 21 base instances mentioned above.

Overview of results. Across all instances, routing and direct allocation account on average for 86% and 14% of the total latency, respectively. Similarly, 85% of customers are served through routes, whereas 15% receive direct service. Interestingly, although approximately 67% of customers lie within the coverage range of at least one depot, only a fraction of them is ultimately served through direct allocation, indicating that coverage alone does not guarantee the use of direct service.

Fig. 5 illustrates the impact of R and F_A on the optimal latency. As expected, increasing the coverage range reduces latency, whereas increasing the direct allocation factor produces the opposite effect. However, the influence of both parameters is highly non-linear.

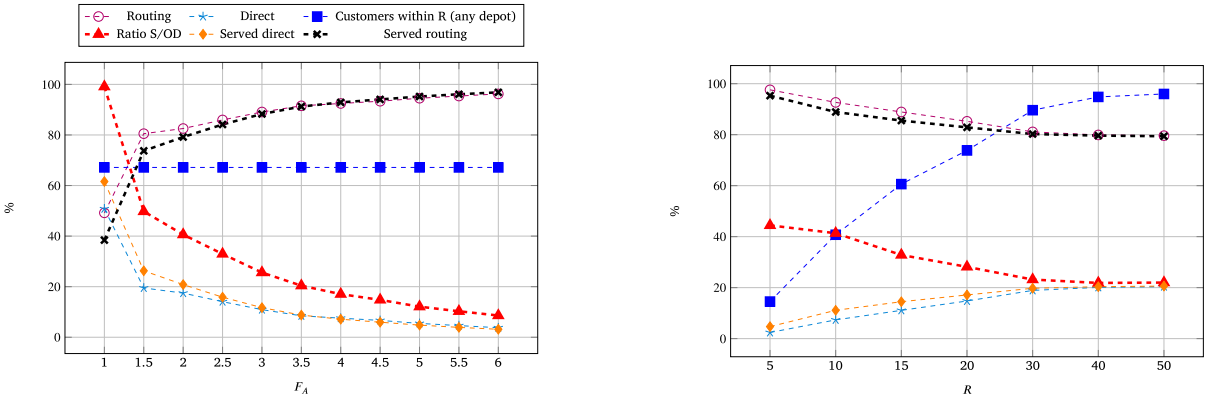
The benefits obtained from increasing the coverage range rapidly diminish as R increases. The average latency reduction is 3.2% when moving from $R = 5$ to $R = 10$, 1.7% from $R = 10$ to $R = 15$, 1.1% from $R = 15$ to $R = 20$, and only 0.8% from $R = 20$ to $R = 30$. Beyond $R = 30$, additional improvements are negligible. This saturation effect is also visible in Fig. 5(b), where the curves corresponding to large values of R progressively overlap.

The effect of F_A follows a similar decreasing pattern. The average latency increase associated with a 0.5-unit increment in F_A decreases from 7.3% when moving from $F_A = 1$ to $F_A = 1.5$ to only 0.3% when moving from $F_A = 5.5$ to $F_A = 6$. Unlike the effect of R, the curves do not fully converge (Fig. 5(a)), because even when the same customers are served directly, larger values of F_A increase their contribution to the objective function. Nevertheless, the increasingly small differences observed for large values of F_A indicate that the solutions rely progressively less on direct service and become predominantly routing-based.

Metrics. To better understand how R and F_A affect the structure of the optimal solutions, Fig. 6 reports several performance metrics: the percentage of the contribution of each mode to the objective function ('Routing', 'Direct'), the percentage of customers served by each such mode ('Served routing', 'Served direct'), the ratio between the number of customers served by direct allocation and the number of customers within the coverage range of any open depot ('Ratio S/OD'), and the percentage of customers within the coverage range of any depot (opened or not). Note that the last metric regards an instance feature, and thus, it is not influenced by the decisions to make, i.e., it depends only on the value of R.

Figs. 6(a) and 6(b) show that the contribution of each service mode closely follows the proportion of customers served through that mode. As expected, the share of customers served directly increases with R and decreases with F_A.

For $R = 5$, only 15% of the customers are, on average, within the coverage range of a depot, and 5% of the customers are served directly. As R increases, the share of customers within coverage grows rapidly, exceeding 90% for $R \geq 30$. However, the proportion served directly changes very little beyond this point, explaining the progressive overlap in the dots depicted in Fig. 5(b). This result suggests that most of the benefits of expanding coverage are already captured at moderate values of R.



(a) For different values of F_A.

(b) For different values of R.

Fig. 6. Metrics associated with optimal solutions.

Table 5
Metrics obtained for the different variants of the LLoRP.

	$\|C\| < 50$					$\|C\| \in [50, 100[$					$\|C\| = 100$			
Metric	1-LLoRP	LLoRP	U-LLoRP	MD	U-MD	1-LLoRP	LLoRP	U-LLoRP	MD	U-MD	LLoRP	U-LLoRP	MD	U-MD
Latency	1597.5	1255.5	1251.6	1170.1	1168.3	1480.7	1084.3	1076.1	960.4	954.8	1612.9	1604.7	1334.8	1333.2
Routing (%)	91.9	83.7	84.2	80.1	80.2	94.7	89.2	89.8	83.2	83.7	85.2	85.9	74.0	74.2
Direct (%)	8.1	16.3	15.8	19.9	19.8	5.3	10.8	10.2	16.8	16.3	14.8	14.1	26.0	25.8
Served Routing (%)	90.5	82.7	83.1	78.1	78.2	91.9	86.3	87.2	80.2	80.8	84.2	84.9	72.9	73.1
Served Direct (%)	9.5	17.3	16.9	21.9	21.8	8.1	13.7	12.8	19.8	19.2	15.8	15.1	27.1	26.9

Fig. 6(a) further shows that when $F_A = 1$, all customers within the coverage range of an open depot are served directly (Ratio S/OD = 100%), in accordance with the properties discussed in Section 4. As F_A increases, direct service becomes less attractive and the Ratio S/OD decreases accordingly. Nevertheless, direct allocation never disappears from the optimal solutions. Even for $F_A = 6$, direct service still accounts on average for 3.15% of customers, corresponding to 8.6% of the customers within the coverage range of an open depot.

The results provide two main insights for practitioners. First, increasing the coverage range can substantially improve responsiveness when the initial coverage is limited, but the marginal gains quickly vanish. Therefore, investments in technologies that extend coverage (e.g., longer-range drones) should be evaluated carefully, as most of the achievable benefit is already captured at moderate values of R. Second, direct service remains a valuable component of the system even under unfavorable access conditions. Hence, decision-makers should not discard this service mode a priori; rather, the results support adopting a hybrid strategy that combines routing and direct allocation across a wide range of operating conditions.

An illustrative example presenting the effect of parameters F_A and R over the optimal decisions of the problem, i.e., facility and transportation mode selection, is presented in Section A1 in the supplementary materials.

6.3.3. Relaxing the capacity constraints of the vehicles and changing the maximum number of depots to be open

In many practical applications, organizations may invest in larger vehicles (i.e., vehicles with effectively unbounded capacity) and increase the number of available operating facilities. Such investments are expected to enhance system performance and reduce customer waiting times. This section aims to address the following managerial questions: *Is it worthwhile to invest in larger vehicles and additional facilities? To what extent can system performance be improved through these investments?* To answer these questions, we conduct four additional sets of computational experiments.

- U-LLoRP: the capacity constraints of the vehicles are removed. This variant can be considered as an uncapacitated LLoRP.
- MD: the constraint related to the maximum number of depots that can be selected is relaxed. In other words, we set $p = |D|$ (considering capacitated vehicles). This variant can be considered a multi-depot cumulative capacitated vehicle routing problem with depot coverage.
- U-MD: relaxes both assumptions, i.e., we assume uncapacitated vehicles and no limit in the number of depots to select. This variant can be considered a multi-depot k-traveling repairman problem with depot coverage, and it corresponds to LB2.
- 1-LLoRP: a single depot can be opened ($p = 1$). The solution of this problem can be considered a cumulative capacitated vehicle routing problem with depot coverage.

Table 5 summarizes the global average results obtained for each variant of the problem. The results are grouped by the size of the instances as done before ($|C| < 50$, $|C| \in [50, 100[$, $|C| = 100$), and the problems are listed from the most to the least constrained

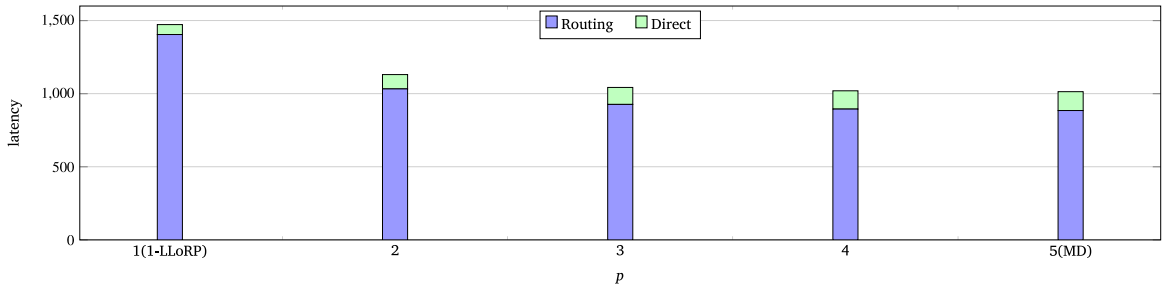


Fig. 7. Average latency for different values of p: instances with $|C| \leq 50$ and $|D| = 5$.

one. Note that in the fourth set of experiments (1-LLoRP) Gurobi runs out of memory for the 2 base instances with $|C| = 100$. For this reason, the column related to 1-LLoRP for $|C| = 100$ is not reported. Each row presents a metric already introduced in Section 6.3.2.

The effect of uncapacitated vehicles. According to the results, relaxing the capacity of the vehicles does not have a significant impact on the solutions, either in terms of objective value or the decisions made. This conclusion is based on two comparisons: LLoRP versus U-LLoRP, and MD versus U-MD. The average latencies obtained with U-LLoRP and U-MD are 0.53% and 0.33% smaller than the average latencies obtained with LLoRP and MD, respectively. Furthermore, comparing the capacitated problem with its uncapacitated version, the average absolute value of the difference of the values for all the remaining metrics is less than 1% (in other words, the structure of the solutions is similar). Analyzing the detailed results we found the following. For LLoRP and U-LLoRP, in 51% of the cases, the optimal solution values of both problems are equal, while for the remaining 49%, the latency can be reduced by 1.38% when uncapacitated vehicles are considered. The savings in latency are mainly derived from more efficient routes; actually, for 76% of the instances, the direct allocation component of the solutions remains unchanged. Similarly, for MD and U-MD, in 64% of the cases the optimal solution values of both problems are equal, while for the remaining 36% the latency can be reduced by 1.39% considering uncapacitated vehicles. For 83% of the instances, the direct allocation component of the solutions remains unchanged and for both problems (MD and U-MD) the optimal solution of all the instances makes use of all the available depots.

From a practical perspective, having uncapacitated vehicles is equivalent to having vehicles with very large capacities. This type of vehicles are usually more expensive than those with *regular* capacities. The main conclusion is that the latency reduction of the system is minor when such an additional investment is made.

The effect of p. According to the results, the maximum number of depots to select has a visible effect on the optimal latency and naturally on the structure of the solutions. Comparing the average optimal latencies associated with each problem (considering all the instances with $|C| <100$) we realize that when $p=1$ the average optimal latency is 31.8% larger w.r.t. that obtained with the original value of p, while when $p = |D|$, the average optimal latency is 9.1% (10.1% when considering all the instances) smaller w.r.t. when the original value of p is considered.

Since for all the instances with $|C| <100$ but one (Christ-75-10), we have $|D|=5$, with $p \in \{2,3\}$, it is worth analyzing in detail the effect of p over the waiting times. Fig. 7 presents the latency associated with the optimal solutions for $p \in \{1,2,3,4,5\}$ for all the instances with $|C| \leq 50$ and $|D| = 5$. The marginal reduction in the latency resulting from selecting an additional depot, computed for each value of $p > 1$ as $100(Latency_{(p-1)} - Latency_p)/Latency_{(p-1)}$, is equal to 23.1%, 7.9%, 2.3%, and 0.4% when two, three, four, and five depots are selected, respectively. This indicates, for example, that selecting two depots instead of one leads to a reduction of 23.1% of the waiting times of the customers. From a decision-maker's point of view, this is an interesting result because the average latency obtained when four depots are selected is almost the same w.r.t. when five depots are selected. Marginal reductions of a considerable magnitude are obtained only until three depots are selected. Accordingly, there is evidence that the number of selected depots has a decreasing marginal contribution to the reduction of the waiting times. Regarding the other metrics presented in Table 5, it is possible to say that the solutions of instances with small values of p are associated with larger latencies and smaller percentages of customers served by direct allocation w.r.t. the solutions of instances with larger values of p.

In post-disaster operations, facilities are typically not constructed from scratch; instead, existing public or private infrastructures, such as schools or stadiums, are repurposed, for instance, as shelters. While this avoids direct construction costs, opening and operating an additional facility still requires resources for setup, staffing, coordination, and ongoing management. Consequently, decision-makers often face budgetary and workforce limitations that prevent the activation of all available facilities. Our results show that, among all the design decisions analyzed, the number of open facilities has the greatest impact on system latency. In particular, the first additional facilities generate substantial reductions in waiting times, whereas the marginal benefit gradually decreases as more facilities are opened. Therefore, determining the appropriate number of facilities to activate is a critical resource-allocation decision, as it allows decision-makers to capture most of the latency improvements while avoiding unnecessary expenditures of personnel and operational resources.

6.3.4. The tradeoff between LLoRP and LoRP

This section investigates the tradeoff between cost and latency in location-or-routing problems. From a managerial perspective, we seek to determine whether decision-makers should prioritize cost, latency, or a compromise between both.

Table 6
Aggregated results obtained when different assumptions are considered.

			Minimizing Cost					Minimizing Latency				
DCC	DFC	RLC	Cost*	Latency	Δ_{Lat^*}(%)	$\lvert K_d \rvert$	$\lvert D_o \rvert$	Cost	Latency*	Δ_{Cost^*}(%)	$\lvert K_d \rvert$	$\lvert D_o \rvert$
x	x	x	143.8	348.6	34.4	1.8	2.4	188.5	241.6	27.2	2.1	2.4
x	x	✓	145.3	345.9	33.2	1.9	2.4	188.6	241.9	26.1	2.1	2.4
x	✓	x	275.7	485.1	109.0	2.2	1.3	375.2	241.6	36.3	2.1	2.3
x	✓	✓	282.4	447.5	99.2	2.4	1.3	375.4	241.9	32.6	2.1	2.3
✓	x	x	148.6	368.2	41.8	1.9	2.4	190.5	246.0	24.1	2.1	2.4
✓	x	✓	150.4	351.0	33.2	2.0	2.4	190.6	246.4	22.7	2.1	2.4
✓	✓	x	331.5	387.8	54.3	2.0	2.4	384.2	246.0	15.2	2.1	2.4
✓	✓	✓	332.9	376.9	50.1	2.1	2.4	384.3	246.4	14.7	2.1	2.4
Average			226.3	388.9	56.5	2.0	2.1	284.7	244.0	24.9	2.1	2.4

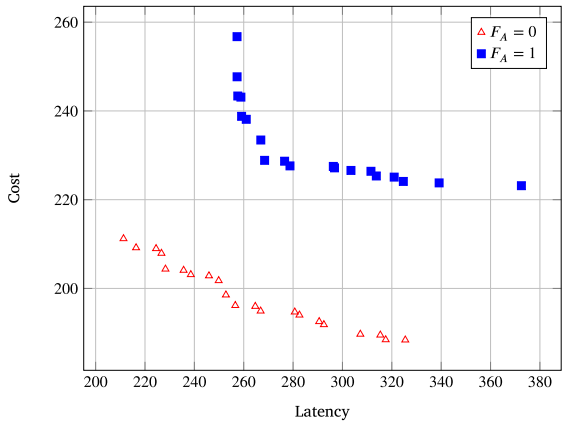
We compare the LLoRP proposed in this paper with the LoRP of Arslan (2021), whose additional features are: (*i*) depot-capacity constraints (DCC), (*ii*) route-length constraints (RLC), and (*iii*) depot fixed costs (DFC). The analysis has two components. First, we evaluate the impact of these features on cost- and latency-optimal solutions. Second, we examine the cost–latency tradeoff through Pareto fronts obtained by our implementation of the ϵ-constrained method.

For the first analysis, the single-objective 2-IF model (Section 5) is solved to optimality under both objective functions and for the eight possible combinations of DCC, RLC, and DFC. For each solution we record the total cost, latency, deviations from the minimum cost and minimum latency (Δ_{Cost^*} and Δ_{Lat^*}), the number of vehicles used ($|K_d|$), and the number of open depots ($|D_o|$). Because route reversals produce identical costs but different latency values, we report the extreme points of the payoff table to avoid overstating the tradeoff. Following Arslan (2021), we consider $R \in \{5, 10, 15, 20, 30, 40, 50\}$, $F_R = 1$, $F_A \in \{0, 0.2, 1\}$, and $T = 100$. Parameters ρ_i and κ_i are taken from the original datasets (Prins et al., 2006; Barreto, 2004). Five benchmark instances (20-5-1, 20-5-1b, 20-5-2, 20-5-2b, and Gaskell-21-5) were selected because all parameter configurations could be solved to proven optimality, yielding 105 test instances.

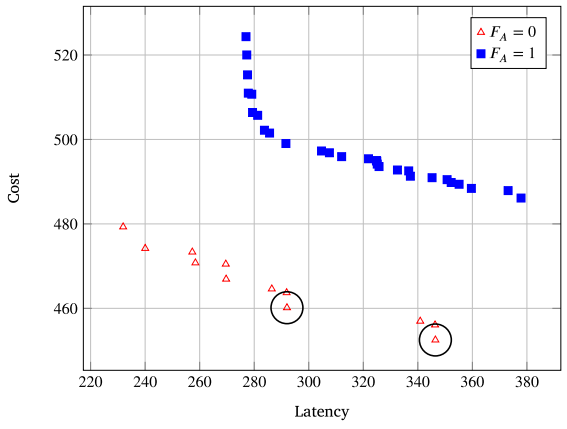
For the Pareto-front analysis, we consider the two extreme settings: the full LoRP (DCC + RLC + DFC) and the LLoRP (none of these features). The ϵ-constrained method is applied with `iter` = 100, using $R = 10$ and $F_A \in \{0, 1\}$. The performance metric is the number of non-dominated solutions (NDS).

The effect of each feature. According to the results presented in Table 6, regardless of the combination of features, there is a tradeoff between both objectives. The average value of Δ_{Lat^*} (Δ_{Cost^*}) indicates that the latency (cost) of the optimal cost-based (latency-based) solutions is on average 56.5% (24.9%) deviated from the optimal latency (cost). From a practical perspective, this is significant; to consider a cost-oriented post-disaster planning whose response time is 56.5% larger than the minimum possible value is equivalent to taking high risks that might lead to human suffering. This is particularly relevant in events like an earthquake or a tsunami, where the evacuation of affected areas must be done as soon as possible. Furthermore, we observe that the values of Δ_{Lat^*} are larger than the respective Δ_{Cost^*} values. This asymmetry indicates that optimal latency-oriented solutions do not increase the cost of the system as much as the cost-oriented solutions increase the latency of the system compared to the best possible cases. The minimum deviation value is obtained when all the assumptions are considered. Nevertheless, it is a considerably large value ($\Delta_{Cost^*} = 14.7\%$). According to the results, the inclusion of the additional LoRP assumptions does not significantly impact the LLoRP solutions: $|K_d|$ and $|D_o|$ remain almost unchanged, while the average latency increases just 2% from the version of the problem with fewer assumptions to the one with the most assumptions. An important observation is that, given the nature of the instances ($F_A \leq F_R$), the optimal solutions of the latency-oriented problem are intensive on direct allocation, leading in some cases to non-utilization of available vehicles. For the LoRP, the omission of the RLC implies optimal solutions with fewer vehicles. Similarly, if DFC are considered, fewer depots are used in the LoRP solutions. Nevertheless, this is noticeable only when the DCC are neglected (due to these constraints, some solutions with few open depots are infeasible). Comparing the optimal solutions obtained under each objective, latency minimization generally results in the use of more vehicles and open depots than cost minimization. An exception arises when DCC are omitted while RLC and DFC are enforced. In this setting, cost-optimal solutions employ more vehicles on average, as fixed depot costs encourage the use of a small number of uncapacitated depots, requiring additional vehicles to satisfy route-length constraints. Finally, since the deviation between both problems is large even when the structure of the solutions is similar (see the row considering all the assumptions), it is possible to say that the differences between both problems also come from the transportation component. In other words, the LoRP and the LLoRP have differences at each decision level: strategic (location decisions), tactical (vehicles fleet), and operational (transportation decisions). Consequently, the decision-maker's strategic emphasis on either cost efficiency or responsiveness plays a critical role in determining system performance.

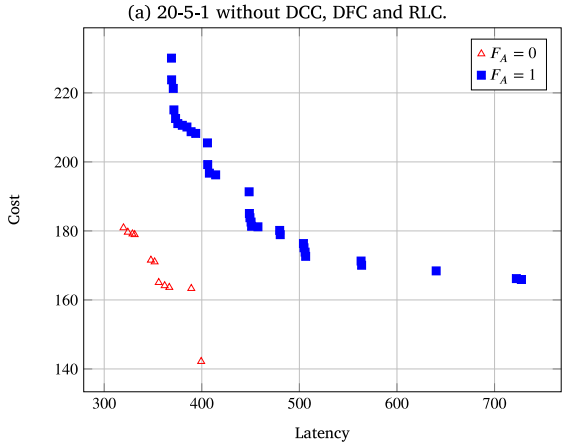
ϵ-constrained method. The bi-objective analysis shows that slightly richer Pareto fronts are obtained when none of the additional LoRP features are considered, with an average of 17.2 non-dominated solutions (NDS) compared to 17.0 when all features are included. The parameter F_A has a stronger influence on the number of NDS. When $F_A = 0$, the average number of NDS is 10.8 for the unconstrained case and 8.2 when DCC, RLC, and DFC are included. In contrast, when $F_A = 1$, these averages increase to 23.6 and 25.8, respectively. This behavior can be explained by the fact that, for $F_A = 0$, customers within the coverage range of an open depot are optimally served



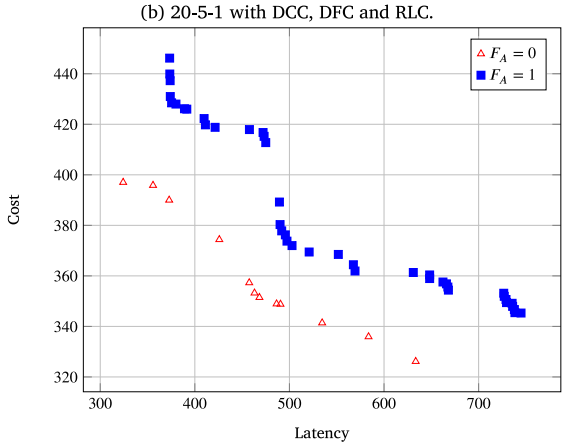
(a) 20-5-1 without DCC, DFC and RLC.



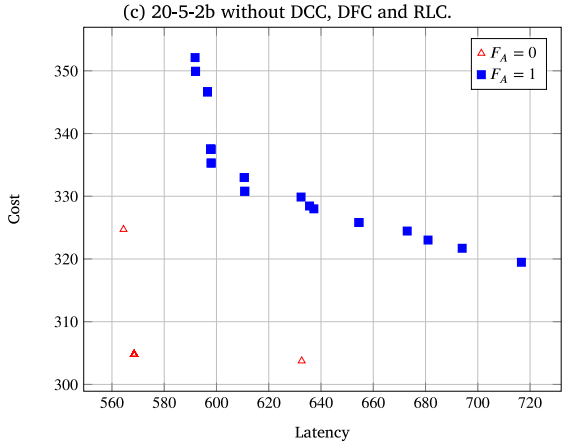
(b) 20-5-1 with DCC, DFC and RLC.



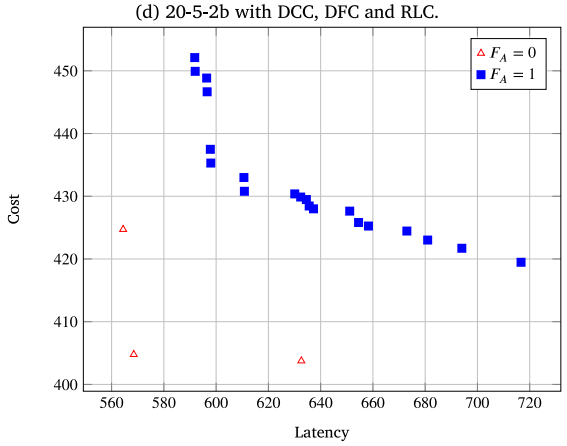
(c) 20-5-2b without DCC, DFC and RLC.



(d) 20-5-2b with DCC, DFC and RLC.



(e) Gaskell-21-5 without DCC, DFC and RLC.



(f) Gaskell-21-5 with DCC, DFC and RLC.

Fig. 8. Pareto fronts for different instances considering $R = 10$.

through direct allocation regardless of the objective considered, reducing the scope for alternative transportation decisions. Fig. 8 depicts the Pareto fronts found for three of the five base instances. The left-hand panels correspond to the LLoRP setting (without DCC, RLC, and DFC), while the right-hand panels correspond to the LoRP setting (with all three features). Across all instances, a clear tradeoff between cost and latency is observed, confirming that this relationship persists regardless of the modeling assumptions adopted.

From a managerial perspective, the availability of Pareto-efficient solutions is preferable to choosing between two extreme solutions. Several fronts reveal situations in which a marginal increase in cost yields substantial reductions in latency (and vice-versa). For example, in Fig. 8(b) with $F_A = 0$, the cost-optimal solution exhibits tradeoffs of $(\Delta_{Lat^*}, \Delta_{Cost^*}) = (49.4, 0)$. Moving to a nearby

Table 7
Managerial takeaways of the computational study.

Section	Managerial question	Main finding	Managerial implication
6.3.1	Is it worth offering direct service in addition to vehicle-based delivery?	Allowing direct service reduces total waiting time by 7.2% on average (up to 19.5%) compared with a routing-only system. Savings remain positive even in unfavorable settings ($R = 5$: 2.1%; $F_A = 6$: 1.4%).	Whenever feasible, include a direct-service option. It improves responsiveness without requiring additional vehicles or facilities and remains beneficial across a wide range of operating conditions.
6.3.2	Is it worth increasing the coverage range R?	Latency reductions exhibit decreasing returns: 3.2% ($R: 5 \rightarrow 10$), 1.7% ($R: 10 \rightarrow 15$), 1.1% ($R: 15 \rightarrow 20$), 0.8% ($R: 20 \rightarrow 30$), and approximately 0% beyond.	Moderate coverage ranges capture most of the attainable benefits. Extending coverage further (e.g., through longer-range drones) is unlikely to provide proportional improvements.
6.3.2	Does direct service remain valuable under unfavorable access conditions?	Latency increases as F_A grows, but at a decreasing rate; direct service remains part of the optimal solution even for $F_A = 6$.	Do not discard direct service solely because access is slow or difficult. Hybrid designs remain beneficial even under adverse accessibility conditions.
6.3.3	Is it worth opening additional facilities?	Opening a second facility reduces latency by 23.1%, a third by 7.9%, a fourth by 2.3%, and a fifth by only 0.4%. This is the most influential design decision analyzed.	Prioritize facility activation when resources are available. In the studied networks, opening two to three facilities captures most of the achievable responsiveness gains.
6.3.3	Is it worth investing in larger vehicles?	Removing vehicle-capacity limits improves latency by less than 1% on average, while the structure of the solutions remains almost unchanged.	Increasing vehicle capacity is generally not an effective strategy for improving responsiveness. Resources are better allocated to facilities or coverage expansion.
6.3.4	Should decision-makers focus on cost or latency?	Cost-optimal plans exhibit latencies 56.5% above the minimum, whereas latency-optimal plans cost 24.9% above the minimum. Pareto fronts reveal attractive intermediate solutions.	Avoid optimizing a single objective. Solutions near the knee of the Pareto front provide substantial reductions in waiting times at a limited additional cost.

Pareto-efficient solution changes these values to (25.9, 1.7), reducing the latency deviation (w.r.t. the optimal latency) by more than 23 percentage points while increasing cost by less than 2% relative to the minimum-cost value. This highlights the value of considering balanced solutions rather than focusing exclusively on either objective.

There are other ways in which the tradeoff between both objectives can be enhanced, e.g., by considering a larger available fleet and by considering that the number of depots to select is a decision to make as in the LoRP. Given their properties and the results analyzed in the first part of this section, the latency-based objective would seek to use as many vehicles and depots as possible, while the cost-based would try to reduce them to an efficient (or necessary) amount. Thus, our goal is not to induce the tradeoff between both objectives by conveniently setting the value of the parameters, but to highlight how they interact under realistic settings. Furthermore, more sophisticated techniques can be employed for obtaining even richer Pareto fronts, which is beyond the scope of this paper (see, e.g., Mavrotas 2009). In our opinion, the results presented in this section open a clear research direction.

Table 7 condenses the practical content of this study. Its guidelines are directly actionable in the applications motivating the LLoRP: a relief-operations planner learns that activating a second and possibly a third distribution hub, with a moderate walking-access range, dominates any investment in larger vehicles or in extended reach, and that enforcing a purely cost-efficient plan would inflate the waiting time of the affected population by more than 50% on average; a retailer learns that offering pickup to nearby customers reduces the waiting time of *all* customers–including those served at home–and that near-cost-optimal operating points with substantially better service levels exist and can be computed exactly with the models proposed in this paper.

7. Conclusions

In this paper, the latency location-or-routing problem (LLoRP) was introduced—a variant of the location-or-routing problem (LoRP) in which the goal is to minimize the total waiting time of the customers. The latter can be served by direct allocation (and not as part of a route) if they lie within a given range of an operating depot. Several combinations of assumptions were considered that unify the existing knowledge for both the LLRP and LoRP. Three compact MILP models were derived. The newly proposed models were adapted to the cost-based LoRP and compared to the state-of-the-art methods. Moreover, two lower bounds were proposed for the new problem. A bi-objective modeling framework was considered to find tradeoff solutions between cost and latency. In particular, strong evidence was provided that those objectives may easily be conflicting, thus calling for such a tradeoff to be sought. An extensive computational study that includes managerial analysis was conducted.

Based upon the extensive results obtained, several relevant conclusions could be drawn: (i) when solving the LLoRP, the multi-layer formulation dominates the two- and the three-index models proposed. Indeed, using the former, it was possible to solve all the tested instances with up to 50 customers, and several instances with up to 100 customers to proven optimality within a computational time limit of two hours. (ii) Clear evidence was gathered that by considering direct allocation in addition to routing, it is possible to significantly decrease the latency, thus giving much relevance to the new investigated problem. (iii) A significant impact can

be observed on the solutions' value and structure due to embedding the direct-allocation cost factor in the models and also due to the range that allows a customer to be directly served from a depot. (iv) The conflicting nature of latency and cost in location-or-routing problems is clear. In particular, the conducted multicriteria analysis evidenced rich Pareto fronts due to the differences at each decision level that latency-based and cost-based solutions have. (v) Regarding the standard LoRP, the results revealed that the two-index formulation is the most effective compact model for solving the problem. In particular, a considerable number of instances with up to 50 customers were solved to proven optimality using the proposed formulations with a time limit of three hours. (vi) Several new best-known solutions could be obtained for LoRP benchmark instances.

A natural extension of this work regards the inclusion of uncertainty in the parameters, namely, in the travel times. Given the customer-centric perspective usually underlying latency-oriented transportation problems, risk-aversion is of clear relevance in such a case. Another interesting direction to follow regards the preferences of the customers. In the investigated LLoRP, the decision maker can decide whether a certain customer is visited by a vehicle or served by direct allocation. Nevertheless, when the direct allocation cost factor is considerably larger than the routing one, the customers might prefer to be served as part of a route. Finally, the exploitation of more sophisticated methods for an in-depth multi-objective analysis defines another potentially interesting extension of this research. In particular, the complexity of the problems being investigated calls for the development of specially tailored approximate procedures for finding Pareto solutions and thus to approximate the Pareto front.

CRediT authorship contribution statement

Alan Osorio-Mora: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Eduardo Álvarez-Miranda:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Francisco Saldanha-da-Gama:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no potential conflict of interests.

Acknowledgments

The authors thank the three anonymous reviewers for the insightful comments and suggestions provided, which helped to improve the manuscript. E. Álvarez-Miranda acknowledges the support of the Complex Engineering Systems Institute, and acknowledges the support of the National Agency of Research and Development (ANID, Chile) through the grant FONDECYT N.1260465 and through the Institute of Ecology and Biodiversity (ANID/BASAL FB210006).

Data availability

Data will be made available on request.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.tre.2026.105280.

References

Akca, Z., Berger, R.T., Ralphs, T.K., 2009. A branch-and-price algorithm for combined location and routing problems under capacity restrictions. In: Operations Research and Cyber-Infrastructure. Springer, pp. 309–330.

Arslan, O., 2021. The location-or-routing problem. Transp. Res. B Methodol. 147, 1–21.

Arslan, O., Çulhan Kumcu, G., Kara, B.Y., Laporte, G., 2021. The location and location-routing problem for the refugee camp network design. Transp. Res. B Methodol. 143, 201–220.

Barreto, S., 2004. Analysis and modelling of location-routing problems. Ph.D. thesis. University of Aveiro. Portugal.

Bayram, V., Kara, B.Y., Saldanha-da-Gama, F., Yaman, H., 2023. Humanitarian logistics under uncertainty: planning for sheltering and evacuation. In: Eiselt, H.A., Marianov, V. (Eds.), Uncertainty in Facility Location Problems. Springer Nature Switzerland, pp. 81–105.

Bowes, P., 2023. Parcel shipping index 2023. Available at: https://www.pitneybowes.com/content/dam/pitneybowes/us/en/shipping-index/23-mktc-03596-2023_global_parcel_shipping_index_ebook-web.pdf (accessed July 3, 2026).

Bruni, M.E., Khodaparasti, S., Beraldi, P., 2020. The selective minimum latency problem under travel time variability: an application to post-disaster assessment operations. Omega 92, 102154.

Bruni, M.E., Khodaparasti, S., Martínez-Salazar, I., Nucamendi-Guillén, S., 2022. The multi-depot k-traveling repairman problem. Optim. Lett. 16 (9), 2681–2709.

Bruni, M.E., Khodaparasti, S., Perboli, G., 2023. The drone latency location routing problem under uncertainty. Transp. Res. C Emerg. Technol. 156, 104322.

Cavagnini, R., Santini, A., Schneider, M., 2025. Recent developments in location-routing problems: deterministic, single-echelon, single-objective, single-period problems. Eur. J. Oper. Res. 332, 711–729.

Corona-Gutiérrez, K., Nucamendi-Guillén, S., Lalla-Ruiz, E., 2022. Vehicle routing with cumulative objectives: a state of the art and analysis. Comput. Ind. Eng. 169, 108054.

Dönmez, Z., Saldanha-da Gama, F., Karsu, O., Bahar, Y.K., Ayyıldız, M., Uslu, B., 2025. Humanitarian logistics: how fair is fairness? Working Paper Series. Social Science Research Network (SSRN).

De la Fuente, R., Aguayo, M.M., Contreras-Bolton, C., 2024. An optimization-based approach for an integrated forest fire monitoring system with multiple technologies and surveillance drones. Eur. J. Oper. Res. 313 (2), 435–451.
eMarketer, 2024. Worldwide ecommerce sales to break $6 trillion. Available at: https://www.emarketer.com/content/worldwide-ecommerce-sales-break-6-trillion (accessed July 3, 2026).
Forbes, 2023. Trending now: Last-mile delivery expectations for 2023. Available at: https://www.forbes.com/councils/forbestechcouncil/2023/02/24/trending-now-last-mile-delivery-expectations-for-2023/ (accessed July 3, 2026).
Grand View Research, 2025. Online food delivery market (2025–2030). Available at: https://www.grandviewresearch.com/industry-analysis/online-food-delivery-market-report (accessed July 3, 2026).
Haghi, M., Arslan, O., Laporte, G., 2023. A location-or-routing problem with partial and decaying coverage. Comput. Oper. Res. 149, 106041.
Lalla-Ruiz, E., Voß, S., 2020. A POPMUSIC approach for the multi-depot cumulative capacitated vehicle routing problem. Optim. Lett. 14 (3), 671–691.
Laporte, G., Nickel, S., Saldanha-da-Gama, F. (Eds.), 2019. Location Science. Springer International Publishing. 2nd edition.
Mara, S. T.W., Kuo, R.J., Asih, A. M.S., 2021. Location-routing problem: a classification of recent research. Int. Trans. Oper. Res. 28 (6), 2941–2983.
Mavrotas, G., 2009. Effective implementation of the ε-constraint method in multi-objective mathematical programming problems. Appl. Math. Comput. 213, 455–465.
Molina, J., López-Sánchez, A.D., Hernández-Díaz, A.G., Martínez-Salazar, I., 2018. A multi-start algorithm with intelligent neighborhood selection for solving multi-objective humanitarian vehicle routing problems. J. Heuristics 24, 111–133.
Moshref-Javadi, M., Lee, S., 2016. The latency location-routing problem. Eur. J. Oper. Res. 255, 604–619.
Nasir, J.A., Kuo, Y.-H., 2024. Stochastic home care transportation with dynamically prioritized patients: an integrated facility location, fleet sizing, and routing approach. Transp. Res. B Methodol. 184, 102949.
Ngueveu, S.U., Prins, C., Calvo, R.W., 2010. An effective memetic algorithm for the cumulative capacitated vehicle routing problem. Comput. Oper. Res. 37, 1877–1885.
Nucamendi-Guillén, S., Gómez-Rocha, J.E., Camacho-Vallejo, J.-F., Pardo, J. A.M., González-Ramírez, R.G., 2026. The latency location routing problem with split deliveries: mathematical formulation and metaheuristic algorithm. Int. Trans. Oper. Res. 33, 1609–1635.
Nucamendi-Guillén, S., Martínez-Salazar, I., Khodaparasti, S., Bruni, M.E., 2022. New formulations and solution approaches for the latency location routing problem. Comput. Oper. Res. 143, 105767.
Osorio-Mora, A., Escobar, J.W., Toth, P., 2023a. An iterated local search algorithm for latency vehicle routing problems with multiple depots. Comput. Oper. Res. 158, 106293.
Osorio-Mora, A., Saldanha-da Gama, F., Toth, P., 2025. A risk-averse latency location-routing problem with stochastic travel times. Eur. J. Oper. Res. 321 (3), 837–850.
Osorio-Mora, A., Rey, C., Toth, P., Vigo, D., 2023b. Effective metaheuristics for the latency location routing problem. Int. Trans. Oper. Res. 30, 3801–3832.
Prins, C., Prodhon, C., Wolfler Calvo, R., 2006. Solving the capacitated location-routing problem by a GRASP complemented by a learning process and a path relinking. 4OR 4, 221–238.
Schneider, M., Drexl, M., 2017. A survey of the standard location-routing problem. Ann. Oper. Res. 259, 389–414.
Segredo, E., Nucamendi-Guillén, S., Miranda, G., Lalla-Ruiz, E., 2026. A matheuristic augmented ε-constraint framework for a bi-objective multi-depot cumulative capacitated vehicle routing problem. Eur. J. Oper. Res.
Sphere Association, 2018. The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response. Geneva. 4th edition.
Sze, J.F., Salhi, S., Wassan, N., 2017. The cumulative capacitated vehicle routing problem with min-sum and min-max objectives: an effective hybridisation of adaptive variable neighbourhood search and large neighbourhood search. Transp. Res. B Methodol. 101, 162–184.
UNWFP, 2016. Nepal Lessons Learned Report: January 2016. Operational Report. Global Logistics Cluster. Rome, Italy. https://logcluster.org/.
UNWFP, 2026. Annual performance report 2025. Available at: https://www.wfp.org/publications/annual-performance-report (accessed July 3, 2026).
Voxware, 2019. Survey: Consumers less tolerant of late and inaccurate deliveries. Available at: https://www.sdcexec.com/warehousing/retail/news/22565436/voxware-inc-consumers-less-tolerant-for-delivery-delays-inaccurate-order-fulfillment-study (accessed July 3, 2026).
Wang, S., Sun, W., Baldacci, R., Elomri, A., 2025. Exact solution of location–routing problems with heterogeneous fleet and weight-based carbon emissions. Transp. Res. E Logist. Transp. Rev. 193, 103862.
Wolfinger, D., Gansterer, M., Doerner, K.F., Popper, N., 2023. A large neighbourhood search metaheuristic for the contagious disease testing problem. Eur. J. Oper. Res. 304 (1), 169–182.
World Economic Forum, 2024. Transforming urban logistics: Sustainable and efficient last-mile delivery in cities. Available at: https://www.weforum.org/publications/transforming-urban-logistics-a-path-to-sustainable-and-efficient-last-mile-delivery-in-cities/ (accessed July 3, 2026).
Zhou, L., Baldacci, R., Mohammed, A.R., 2026. Location-routing problem for robot deliveries with customer choices and hybrid facilities. Transp. Res. E Logist. Transp. Rev. 212, 104889.
Zou, Y., Hao, J.-K., Wu, Q., 2024. A reinforcement learning guided hybrid evolutionary algorithm for the latency location routing problem. Comput. Oper. Res. 170, 106758.