Fair Vehicle Routing via Bilevel Optimization

Ivana Ljubić
ESSEC Business School
France
ljubic@essec.edu

Justo Puerto
Department of Statistics and Operational Research
Institute of Mathematics of the University of Seville
University of Seville, Spain
puerto@us.es

Alberto Torrejon
Department of Statistics and Operational Research
Institute of Mathematics of the University of Seville
University of Seville, Spain
atorrejon@us.es

March 5, 2026

Abstract

We propose a novel approach to modeling fairness in the Vehicle Routing Problem (VRP) by introducing objective functions based on ordering route lengths, capturing both monotonic and non-monotonic equity measures. Our method ensures allocations that are efficient, capacity-feasible, and equitable according to criteria like min-max, range, Gini, variance, or absolute deviations. To prevent biased or inefficient routes, we enforce per-vehicle TSP-optimality, making each route the shortest for its assigned customers. This creates a combinatorial dependency between routing and assignment, modeled via a bilevel optimization framework. We develop mixed-integer linear programming formulations embedding TSP subproblems using value function reformulations, allowing exact solutions. Computational experiments show that enforcing TSP-optimality is crucial for equitable routing under non-monotonic objectives and highlight the structural impact of different fairness criteria.

Keywords Bilevel Optimization · Fairness · Ordered Optimization · Vehicle Routing

1 Introduction

Traditional optimization models primarily emphasize *efficiency*, typically defined in terms of minimizing a global cost function. However, as noted by Savas (1978), focusing on efficiency may be insufficient when addressing, e.g., public administration decision-making, where all segments of the population must receive uniform or non-discriminatory treatment. It is therefore necessary to incorporate a second objective into the decision-making process: *fairness*. Rooted in Equity Theory (Adams, 1963), fairness draws attention to the perceived equity of outcomes. A solution that performs well at the system level but imposes disproportionate burdens on specific groups may be socially rejected or prove unsustainable. Despite its intuitive appeal, fairness remains an elusive and multifaceted concept, with no universally accepted definition. Consequently, the choice of a fairness metric can significantly influence the interpretation and quality of a given solution.

The literature has proposed a myriad of fairness approaches, each designed to capture a specific aspect of equity, such as proportionality, envy-freeness, or balance, but yet these metrics are often mutually incompatible and insufficient to simultaneously guarantee all desirable fairness properties. Furthermore, the notion of fairness entails intrinsic limitations that are, by nature, unavoidable. For example, one of the oldest and most accepted approaches to equity, introduced by Rawls (1971), consists in focusing on the worst-off entity. Some studies aim to maximize this worst outcome, while others impose a constraint to ensure it is below a predefined value. However, this concept has faced substantial criticism, as prioritizing only the least advantaged can lead to globally inefficient solutions, potentially resulting in significant aggregate losses for the remaining agents. Also, there exists an inherent trade-off between fairness and efficiency broadly discussed in the literature, commonly referred to as the *price of fairness*, see Mandell (1991) or Bertsimas et al. (2011). This notion quantifies the relative efficiency loss incurred when fairness considerations are imposed, highlighting that achieving a more balanced distribution of resources often requires that some agents accept suboptimal outcomes to those they would receive under a purely efficiency-oriented scenario.

Mathematical optimization has traditionally served as the cornerstone of decision-making problems with its focus on maximizing efficiency. However, this perspective often neglects the critical dimension of distributive justice, wherein the allocation of costs and benefits among individual agents is a paramount. For this reason, a significant body of literature shifted the focus, advocating for the integration of fairness as a primary objective, compelling decision-makers to navigate the complex Pareto frontier between system efficiency and equitable distribution (see Hooker and Williams, 2012; Bertsimas et al., 2012, among others). One of the most extensively studied family of decision-making problems in the mathematical optimization literature serves as an ideal testbed for exploring this dichotomy: *vehicle routing problems*.

The Vehicle Routing Problem (VRP), first introduced by Dantzig and Ramser (1959), has long stood as a central topic in Operations Research and continues to grow in significance within the increasingly logistics-oriented landscape of contemporary courier companies, supply chains necessities, etc. (see, e.g., Archetti et al., 2025). Despite its simple formulation, the VRP is NP-hard, as it generalizes the *Traveling Salesman Problem* (TSP). Whereas the TSP seeks the shortest possible tour for a single vehicle that visits each customer exactly once and returns to the depot, the VRP extends this framework to multiple vehicles subject to capacity constraints, serving the given set of customers. Current advances in VRP have been largely driven by mathematical programming techniques, most notably branch-and-price algorithms grounded in the set-partitioning formulation of Balinski and Quandt (1964). Among the leading exact solvers developed along these lines is the *VRPSolver* (see Pessoa et al., 2020; Errami et al., 2024). For comprehensive surveys covering both exact and heuristic algorithms, see Laporte (2009), Toth and Vigo (2014), or Archetti et al. (2025), among others.

While effective in optimizing global objectives like total travel distance or the demand served, the resulting VRP solutions can exhibit severe imbalances in driver workload. These disparities can manifest in route duration, distance traveled, or service requirements, leading to decreased driver satisfaction, increased employee turnover, and long-term degradation of service quality, potentially fostering perceptions of unfairness, or even envy, among drivers. Consequently, embedding fairness within vehicle routing models is not merely an ethical consideration but a strategic imperative for achieving robust and sustainable logistics operations. To address this, alternative objective functions that are less sensitive to extreme values have been proposed. These include robust formulations and fairness measures aimed to promote more equitable route distributions. For example, balance can be enforced by minimizing differences between route lengths or explicitly addressing *envy-freeness*. Envy-freeness requires that no individual perceives another's allocation as more favorable. In the context of routing, this translates to assigning routes such that no driver would prefer another's route over their own. In practice, this concept is known to be partially captured by means of the well-known Gini coefficient (Gini, 1921), see, e.g., Blanco et al. (2024) for a successful application in the facility location context. Other common fairness measures in the literature, include min-max, range (difference between maximum and minimum), the variance/standard deviation or the absolute deviation with respect to the mean. The interested reader is referred to Marsh and Schilling, 1994; Karsu and Morton, 2015; Lozano et al., 2016; Matl et al., 2018, among others, for extensive catalogues of fairness measures and detailed analyses of their properties.

A comprehensive and systematic study of fairness in routing, often termed as *workload equity*, was conducted by Matl et al. (2018). They provide a formal taxonomy, deconstructing fairness concepts into (1) the *fairness metrics*, which defines the aspect of workload to be equalized (e.g., route length, service time, or demand served); (2) the *fairness measure*, which quantifies inequality across agents (e.g., range, variance, maximum); and (3) the *fairness objective*, the combination of a metric and a measure, which is the actual function subject to optimization. A critical property they analyze is the *monotonicity* of the fairness objective. An objective is said to be monotonic if a unilateral increase in a single driver's workload cannot improve the overall equity value. Monotonicity is desirable, as it ensures that worsening an individual's condition does not falsely appear beneficial in terms of equity. Examples of monotonic measures include the average (which is not equitable but egalitarian), as well as min–max or k-max measures, which focus on minimizing the k largest costs. However, many widely adopted and intuitive measures of fairness, such as the range, variance, and Gini coefficient, are unfortunately non-monotonic.

Non-monotonicity induces a counterintuitive modeling dilemma, hereinafter termed as *route inconsistency*: an optimization model under a non-monotonic measure may favor a solution where a driver's route is made arbitrarily inefficient, i.e., not a TSP-optimal tour for the set of assigned customers. Here, a route is said to be TSP-optimal if it corresponds to a minimum-length Traveling Salesman tour over the set of customers assigned to the driver. This may occur simply because the resulting increase in workload brings it closer to other routes, spuriously reducing the overall measure of dispersion. Clearly, this is not a desirable property, as non-TSP-optimal tours contradict the minimization of individual workloads. More generally, non-monotonic objectives can be improved even in the extreme case when all workloads are strictly worsened. This does not happen if the objective is monotonic, for which all the solutions are composed only of TSP-optimal tours. To further investigate these pathological behaviors, Matl et al. (2018) introduce the concept of *workload inconsistency*. This phenomenon occurs when a solution in which the individual fairness metrics are strictly worse, therefore incurring a higher total cost, yields a more favorable total value of the equity measure, i.e., a lower level of system inequality. Workload inconsistency exemplifies the paradoxical issue of the price of fairness, whereby improvements in equity require sacrifices in individual efficiency.

Consequently, much of the fairness-oriented literature in vehicle routing focuses on monotonic criteria, predominantly the min-max objective, which has been addressed with exact methods by authors such as Campbell et al. (2008), Huang et al. (2011), and Bertazzi et al. (2015). The trade-off between efficiency and equity has also been explicitly modeled as a bi-objective problem, notably in the *Balanced VRP*, first studied by Jozefowiez et al. (2002) and later advanced with exact methods by Reiter and Gutjahr (2012), Halvorsen-Weare and Savelsbergh (2016), and Artigues et al. (2018). Simultaneously, the challenge of enforcing TSP-optimal routing in multi-depot settings has been studied in the *multiple Traveling Salesman Problem* (mTSP), introduced by Svestka and Huckfeldt (1973). Fairness studies within the mTSP framework have also largely relied on monotonic criteria like the min-max objective (Applegate et al., 2002) or, more recently, ℓ_p norms (Bektaş and Letchford, 2020). Research on non-monotonic fairness objectives is comparatively scarce, with some contributions including an enumeration-based analysis by Matl et al. (2018) and post-hoc evaluations in both heuristic (Lozano et al., 2016) and exact (López Sánchez et al., 2024) contexts. In van Rossum et al. (2025), the authors consider fairness via non-monotonic measures using two specific metrics, range and Gini coefficient. However, they handle TSP-optimality in a heuristic manner, and they conclude, based on an extensive computational study, that embedding TSP-optimality constitutes a challenging task. To the best of our knowledge, our approach is the first work that explicitly considers fairness via non-monotonic measures while simultaneously enforcing TSP-optimal routing with provable optimality guarantees.

Contributions

To address the multifaceted challenges of fairness in vehicle routing, this work introduces several novel contributions to the state-of-the-art. First, we propose a general and flexible modeling framework based on the principles of Ordered Optimization. By ordering route costs, one can directly identify the most and least burdened agents, enabling the adoption of fairness measures that explicitly account for inequality. These order-based measures, originally introduced as L-statistics, trace their theoretical roots to Daniell (1920) and gained practical prominence in robust statistics through the work of Huber (2009), who emphasized their resilience to outliers compared to classical measures like the mean. In decision-making and information aggregation, these measures were later formalized as Ordered Weighted Averaging (OWA) operators by Yager (1988), defining a versatile, parameterized family of mean-type operators widely applied across diverse domains.

Within Operations Research, one of the most significant applications is found in Facility Location Theory, where they are known as Ordered Median Functions, see Nickel and Puerto (2006). While ordered optimization has been extensively studied in location contexts (Boland et al., 2006; Labbé et al., 2017; Deleplanque et al., 2020; Marín et al., 2020; Pozo et al., 2024; Ljubić et al., 2024; Cherkesly et al., 2025), with increasing focus on fairness-oriented formulations (Mesa et al., 2003; Blanco and Gázquez, 2023; Blanco et al., 2024; Tsang and Shehadeh, 2025), its application in routing remains nascent. A more recent and influential development in this field is the introduction of k-sum operators, which generalize classical order-based functions and have attracted substantial interest, see Kalcsics et al. (2002); Ogryczak and Tamir (2003); Mansini et al. (2007); Puerto et al. (2017); Marín et al. (2020); Ljubić et al. (2024); Puerto and Torrejon (2025), among others.

To the best of our knowledge, this paper is the first to introduce the ordered-measure structure to the Vehicle Routing Problem (VRP). Our framework integrates both basic ordered statistics and k-sum approaches to model a broad spectrum of objectives. Furthermore, following the principles in Blanco et al. (2025), we extend this framework to accommodate both linear and non-linear fairness metrics, such as the variance. This paradigm also encompasses fairness criteria based on minimizing absolute deviations from statistical anchors (e.g., the minimum or median workload), providing a unifying structure for equity modeling in routing. Also, our framework allows to use negative entries of the weight vector, which is the case for most of these measures concerning fairness, in particular, non-monotonic ones.

Beyond the modeling framework, our main technical contribution is a novel methodology to address the route inconsistency problem inherent in non-monotonic fairness objectives. We provide for the first time exact approaches that enforce TSP-optimal routing for each individual driver in a provably optimal manner within a fair VRP context. We achieve this through a sophisticated bilevel optimization model. The interested reader is referred to Dempe (2002); Dempe and Zemkoho (2013); Kleinert et al. (2021), among many others, for further details on Bilevel Optimization. In our work, the upper-level problem determines the strategic partitioning of customers among drivers while imposing capacity constraints and optimizing the fairness objective. Simultaneously, the lower-level problem solves a collection of independent TSPs to determine the exact cost of each route. This bilevel architecture ensures a clear separation between the strategic design of fairness and the tactical realization of routing efficiency. Consequently, the trade-offs between fairness and efficiency are made explicit, and operational plans remain unaffected by the distortions typically introduced by non-monotonic objectives.

From a computational perspective, the proposed formulations exhibit markedly different performance depending on both the problem size and the class of fairness objective considered. In particular, we explore two complementary modeling strategies to handle the bilevel enforcement of TSP-optimality. Dense formulations explicitly represent the lower-level routing problems by introducing dedicated routing variables and families of constraints that encode the TSP structure for each driver, yielding strong and accurate models at the expense of the increased model's size. To mitigate this computational burden, we also propose sparse formulations in which routing variables are projected out and route costs are captured implicitly through value-function representations and dynamically generated penalties.

The numerical experiments provide clear evidence that these modeling choices play a crucial role in tractability. Sparse formulations are highly effective for small and medium-sized instances, benefiting from reduced variable counts, whereas dense formulations offer superior robustness and tighter bounds as instance size increases. More generally, the results highlight the intrinsic computational difficulty of non-monotonic fairness objectives. These findings provide valuable insight into the practical trade-offs between model strength, scalability, and fairness expressiveness, and motivate the tailored use of different formulations depending on the operational context.

Structure of the paper

The remainder of this paper is organized as follows. Section 2 introduces the general modeling framework based on the ordering of route lengths. This section demonstrates how to incorporate a broad spectrum of fairness measures into vehicle routing, enabling the modeling of diverse loss functions and the analysis of their structural properties. Two specific formulations are provided: one based on ordering individual route lengths and another leveraging the k-sum operator. Moreover, the framework is also extended to model non-linear measures. Section 3 discusses the critical role of non-monotonicity and TSP-optimality in fairness contexts, establishing the necessity of a bilevel framework to overcome the theoretical and practical challenges inherent in these concepts. Section 4 presents the bilevel optimization model, which integrates the ordered objective framework while enforcing TSP-optimality for each driver's route. In Section 5, the bilevel model is presented as a single-level problem using the value function reformulation, highlighting the structural similarities between the lower-level problem and other TSP variants. Section 6 explores alternative, more parsimonious models where routing variables are projected out, and introduces bounding techniques designed to enhance computational performance. Finally, Section 7 provides an empirical comparison of the different modeling and solution strategies developed in this work, and Section 8 presents concluding remarks and future research directions.

2 Ordered framework for Fair Vehicle Routing

The *Capacitated Vehicle Routing Problem*, which we will refer to indistinctly as VRP, is a classical combinatorial optimization problem that generalizes the well-known *Traveling Salesman Problem*. Its main objective is to determine a set of minimum-cost routes for a fleet of identical vehicles to satisfy the demands of geographically dispersed customers, subject to vehicle capacity constraints. The VRP is defined on a complete directed graph $G = (V, A)$, where $V = \{0, 1, \ldots, n\}$ is a set of nodes and $A = \{(i, j) \in V \times V, i \neq j\}$ is a set of arcs. The node 0 represents the depot for a fleet of vehicles given by the set K with $|K| = m$, where each vehicle has identical capacity $Q > 0$. The set $V' = V \setminus \{0\}$ represents the set of customers that need to be visited. Each customer holds a demand $q_i \geq 0, i \in V'$. It is assumed that node 0 has no demand, i.e., $q_0 = 0$. Every arc $(i, j) \in A$ is associated with the positive travel cost c_{ij}. It is also assumed that the cost matrix satisfies the triangular inequality, $c_{ij} + c_{jl} \geq c_{il}$ for all $i, j, l \in V$. The aim is to determine a minimum-cost set of m routes in order to satisfy the demands of all customers under the following constraints: (1) each route starts and ends at the depot, (2) each customer is visited exactly once, and (3) the total demand of all customers on any route must not exceed the vehicle capacity.

The goal of this work is to address fairness issues in routing. We will consider the VRP problem as a baseline and we will introduce a flexible and unifying framework that accommodates a wide range of fairness and equity measures. To describe the main idea, we use index ℓ to denote the ℓ-th shortest route in a feasible solution, $\ell \in K$. A non-negative variable $\Phi_\ell \geq 0$, will represent the length of the ℓ-th shortest route. By sorting these route lengths in non-decreasing order and associating them with a weight vector $\boldsymbol{\lambda} = (\lambda_\ell)_{\ell=1}^m$, where $\lambda_\ell \in \mathbb{R}$, we obtain a general modeling framework. This framework allows to model various objective functions by changing the weight vector $\boldsymbol{\lambda}$. Table 1 presents several examples of $\boldsymbol{\lambda}$-vectors corresponding to well-known objective functions in the literature, such as mean, min-max or

Statistic	Expression	$\boldsymbol{\lambda}$
Mean	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}(1, ..., 1)$
Maximum	$\max x_i$	$(0, ..., 0, 1)$
Minimum	$\min x_i$	$(1, 0, ..., 0, 0)$
Median	$\text{median} x_i = \begin{cases} x_{(m+1)}, & n = 2m+1 \\ \frac{x_{(m)}+x_{(m+1)}}{2}, & n = 2m \end{cases}$	$(0, ..., 0, \underbrace{1}_{(m+1)-th}, 0, ..., 0)$ $(0, ..., 0, \underbrace{\frac{1}{2}}_{m-th}, \underbrace{\frac{1}{2}}_{(m+1)-th}, 0, ..., 0)$
q-th quantile (or τ-quantile)	$x_{(q)}, q = \lfloor \tau n \rfloor$	$(0, ..., 0, \underbrace{1}_{q-th}, 0, ..., 0)$
Midrange	$\frac{1}{2}(\max x_i + \min x_i)$	$\frac{1}{2}(1, 0, ..., 0, 1)$
General midrange (midhinge, etc.)	$\frac{1}{2}(x_{(r)} + x_{(s)}), (r \leq s)$	$\frac{1}{2}(0, ..., 0, \underbrace{1}_{r-th}, 0, ..., 0, \underbrace{1}_{s-th} ..., 0)$
(r, s)-trimmed mean	$\frac{1}{n-r-s}\sum_{i=r+1}^{n-s} x_{(i)}$	$\frac{1}{n-r-s}(\underbrace{0, ..., 0}_{r}, \underbrace{1, ..., 1}_{n-r-s}, \underbrace{0, ..., 0}_{s})$
(r, s)-mid mean	$\frac{1}{r}\sum_{i=1}^{r} x_{(i)} + \frac{1}{s}\sum_{i=n-s}^{n} x_{(i)}$	$(\underbrace{\frac{1}{r}, ..., \frac{1}{r}}_{r}, \underbrace{0, ..., 0}_{n-r-s}, \underbrace{\frac{1}{s}, ..., \frac{1}{s}}_{s})$
(r, s)-winsorized mean	$\frac{1}{n}\Big(rx_{(r+1)} + \sum_{i=r+1}^{n-s} x_{(i)} + sx_{(n-s)}\Big)$	$\frac{1}{n}(0, ..., 0, \underbrace{r+1}_{(r+1)-th}, 1, ..., 1, \underbrace{s+1}_{(n-s)-th}, 0, ..., 0)$
Range	$\max x_i - \min x_i$	$(-1, 0, ..., 0, 1)$
General interquantile range (e.g., interquartile or interdecile ranges)	$x_{(r)} - x_{(s)}, (r < s)$	$(0, ..., 0, \underbrace{-1}_{r-th}, 0, ..., 0, \underbrace{1}_{s-th} ..., 0)$
Gini differences or mean absolute differences	$\frac{1}{n^2}\sum_{i,j=1}^{n} \lvert x_i - x_j \rvert$	$\frac{2}{n^2}((2i-n-1))_{i \in N}$
Mean absolute deviation from the q-th quantile ($\text{MAD}_{(q)}$)	$\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - x_{(q)} \rvert$	$\frac{1}{n}(-1, ..., -1, \underbrace{2q-n-1}_{q-th}, 1, ..., 1)$

Table 1: Some different objective functions measures and λ-vectors.

quantiles, but also some others which represent more complex modeling choices such as range, Gini or deviations which can be expressed linearly, in the space of variables Φ_ℓ.

From a high-level perspective, we can model the problem as follows:

$$l_{\text{ord}}(\Phi, \boldsymbol{\lambda}) = \min \sum_{\ell \in K} \lambda_\ell \Phi_\ell, \tag{1a}$$

$$\text{s.t. } \Phi_\ell \leq \Phi_{\ell+1}, \qquad \forall \ell \in K : \ell < m, \tag{1b}$$

$$\Phi_\ell \text{ is the length of route } \ell, \qquad \forall \ell \in K, \tag{1c}$$

where in the objective function (1a) we minimize the weighted sum of the variables which are ordered by constraints (1b). The objective function (1a) is commonly referred to in other optimization and operations research contexts as L-*measure* or L-*statistic*, *ordered weighted average operator* (if $\boldsymbol{\lambda} \geq 0$), or *ordered median function*. Constraints (1c) ensure that Φ_ℓ represents the length of the proper route. Note that we are omitting here routing constraints that may depend on the problem under study (such as the number of visits of each client, capacity with respect to demand, capacity with respect to the length of the routes, etc.), so this framework can be used in many others vehicle routing problems.

Regarding the ordering component of our problem, an alternative modeling based on ℓ-sum optimization can be introduced. To the best our knowledge, these concepts have not been studied in the context of routing. Let Ψ_k refer to the length of the route of driver k (before sorting). We now define the following operator as the sum of the largest $m - \ell + 1$ routes, that is:

$$S_\ell(\Psi) = \sum_{i=\ell}^{m} \Psi_{(i)}, \tag{2}$$

where $(\Psi_{(1)}, ..., \Psi_{(m)})$ is the sorted sequence of the route lengths in non-decreasing order, that is, $\Psi_{(1)} \leq ... \leq \Psi_{(m)}$. The reader may notice the relationship between $\Phi_i = \Psi_{(i)}$ provided that all route lengths are different. Using this operator, the expression of the objective function (1a) can be rewritten as:

$$l_{\text{ord}}(\Psi, \lambda) = \sum_{\ell=1}^{m} \lambda_\ell \Psi_{(\ell)} = \sum_{\ell=1}^{m} \Delta_\ell S_\ell(\Psi), \tag{3}$$

where $\lambda_0 := 0$ and $\Delta_\ell := \lambda_\ell - \lambda_{\ell-1}$, for all $\ell \in K$, since we have as possible orders as routes. For further purposes, let us also define the sets $\Delta^- = \{\ell \in K \mid \Delta_\ell < 0\}$, and $\Delta^+ = \{\ell \in K \mid \Delta_\ell > 0\}$. Now, observe that computing (2) is equivalent to solving the following linear problem or its dual (see, e.g., Kalcsics et al., 2002; Ogryczak and Tamir, 2003; Puerto et al., 2017):

$$\begin{aligned} \max & \sum_{k\in K} \theta_k \Psi_k, \\ \text{s.t.} & \sum_{k\in K} \theta_k = m - \ell + 1, \\ & 0 \le \theta_k \le 1, \ \forall k \in K. \end{aligned} \tag{4}$$

$$\begin{aligned} \min\ & (m - \ell + 1)t + \sum_{k\in K} v_k, \\ \text{s.t.}\ & t + v_k \ge \Psi_k, \ \forall k \in K, \\ & t \in \mathbb{R}, \\ & v_k \ge 0, \ \forall k \in K. \end{aligned} \tag{5}$$

Observe that the variables θ_k correspond to the decision variables of a special case of the knapsack problem and, therefore, can be constrained to be binary, i.e., $\theta_k \in \{0, 1\}$, $k \in K$. Moreover, since in our case $\Psi_k \in \mathbb{R}_{\ge 0}$, for all $k \in K$, the variable t can be assumed to be non-negative, i.e., $t \ge 0$.

Given the expression in (2) and the primal-dual relationships inherent to the operator, these elements can be integrated to formulate the following model:

$$\min \sum_{\ell\in\Delta^+} \Delta_\ell \big((m - \ell + 1)t_\ell + \sum_{k\in K} v_{k\ell}\big) + \sum_{\ell\in\Delta^-} \Delta_\ell \sum_{k\in K} \theta_{k\ell}\Psi_k \tag{6a}$$

$$\text{s.t. } t_\ell + v_{k\ell} \ge \Psi_k \qquad k \in K, \ell \in \Delta^+, \tag{6b}$$

$$\sum_{k\in K} \theta_{k\ell} = m - \ell + 1 \qquad \ell \in \Delta^-, \tag{6c}$$

$$t_\ell, v_{k\ell} \ge 0, \qquad k \in K, \ell \in \Delta^+, \tag{6d}$$

$$\theta_{k\ell} \in \{0, 1\}, \qquad k \in K, \ell \in \Delta^-, \tag{6e}$$

$$\Psi_k \text{ is the length of the route of the driver } k, \qquad k \in K. \tag{6f}$$

Note that the objective function (6a) is non-linear because it includes the product of binary variables θ and non-negative Ψ variables. These products can be easily linearized via McCormick envelopes by defining a non-negative continuous variables $\eta_{k\ell} \ge 0$, $k \in K, \ell \in \Delta^-$, representing such products, and adding the constraints:

$$\eta_{k\ell} \le \overline{\Psi}\theta_{k\ell} \qquad k \in K, \ell \in \Delta^-, \tag{7a}$$

$$\eta_{k\ell} \le \Psi_k \qquad k \in K, \ell \in \Delta^-, \tag{7b}$$

$$\eta_{k\ell} \ge \Psi_k - (1 - \theta_{k\ell})\overline{\Psi} \qquad k \in K, \ell \in \Delta^-. \tag{7c}$$

It is important to observe that constraints (7c) are redundant due to the minimization nature of the problem, so they can be removed. Moreover, these constraints involve a big-M constant, $\overline{\Psi}$, which is an upper bound on the route length (see Section 6 for more details).

Furthermore, from the modeling perspective, one could also consider non-linear fairness measures that account for deviation with respect to some reference ordered value representing, e.g., the mean or a quantile of all routes. To this end, we will replace the objective function (1a) with the following loss function:

$$\mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})) = \mathcal{L}(\Psi, l_{\text{ord}}(\Psi, \boldsymbol{\lambda})). \tag{8}$$

For example, if we minimize the *least absolute deviation* with respect to the reference ordered value expressed as $l_{\text{ord}}(\Phi, \boldsymbol{\lambda}) = \sum_{k\in K} \lambda_k \Phi_k$, we have

$$\mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})) = \sum_{k\in K} |\Phi_k - l_{\text{ord}}(\Phi, \boldsymbol{\lambda})|. \tag{9}$$

Similarly, to minimize *the least squared deviation*, we have

$$\mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})) = \sum_{k\in K} (\Phi_k - l_{\text{ord}}(\Phi, \boldsymbol{\lambda}))^2. \tag{10}$$

Nodes	0	1	2	3	4	5	6
Coordinates	(0,10)	(10,0)	(-10,0)	(0,-10)	(7.7,0)	(-7.7,0)	(0,-7.7)
Demands	0	1	1	3	1	1	3

Table 2: Example parameters

		no TSP-optimal				TSP-optimal			
Measure	Monotonicity	Obj	Tour 1	Tour 2	Total	Obj	Tour 1	Tour 2	Total
Mean	yes	47.04	45.03	49.06	94.09	47.04	45.03	49.06	94.09
Median	yes	47.04	45.03	49.06	94.09	47.04	45.03	49.06	94.09
Max	yes	49.06	45.03	49.06	94.09	49.06	45.03	49.06	94.09
Range	no	0.04	56.57	56.61	113.18	3.04	50.48	53.52	104.00
Gini	no	0.02	56.57	56.61	113.18	1.52	50.48	53.52	104.00
$\mathrm{MAD}_{(med)}$	no	0.02	56.57	56.61	113.18	1.52	50.48	53.52	104.00
Variance	no	$4e^{-4}$	56.57	56.61	113.18	2.31	50.48	53.52	104.00

Table 3: Example results for different choices of $\boldsymbol{\lambda}$

If we set

$$\mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})) = l_{\text{ord}}(\Phi, \boldsymbol{\lambda}), \tag{11}$$

we are simply minimizing the initial ordered measure defined by vector $\boldsymbol{\lambda}$.

3 Dealing with non-monotonicity and route inconsistency

For the sake of completeness, we start by reviewing next the concept of monotonicity.

Definition 1 (Monotonicity) *Let $\Phi, \Phi' \in \mathbb{R}^m_{\geq 0}$ be two vectors of route lengths each associated with a feasible solution of the VRP, where $\Phi = (\Phi_k)_{k \in K}$ and $\Phi' = (\Phi'_k)_{k \in K}$. Let the value of the fairness (or inequality) measure be defined by the loss function*

$$\mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})), \tag{12}$$

where $l_{\text{ord}}(\Phi, \boldsymbol{\lambda})$ is the ordered weighted objective introduced in (1a)*.*

We say that $\mathcal{L}(\cdot)$ is monotone if the following condition holds: for any Φ' such that

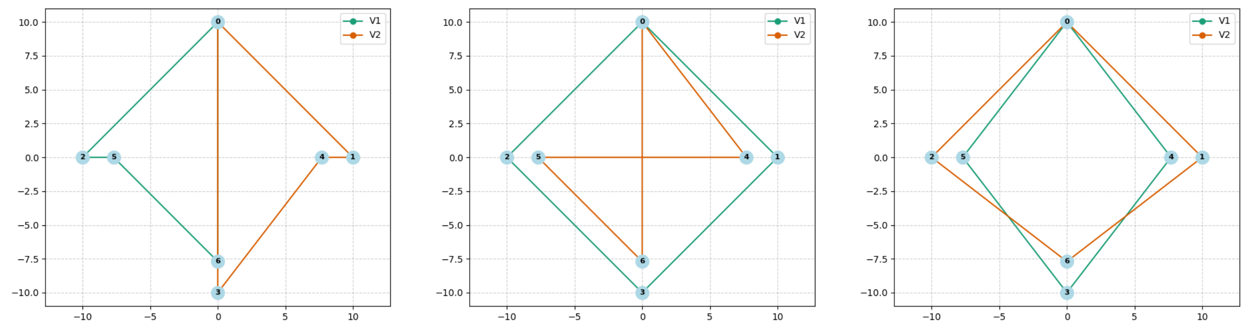
$$\Phi'_k = \Phi_k + \delta_k, \qquad \forall k \in K,$$

with $\delta_k \geq 0$ for all $k \in K$ and $\exists\, k \in K$ with $\delta_k > 0$, it follows that

$$\mathcal{L}(\Phi', l_{\text{ord}}(\Phi', \boldsymbol{\lambda})) \;\geq\; \mathcal{L}(\Phi, l_{\text{ord}}(\Phi, \boldsymbol{\lambda})).$$

In other words, worsening at least one route while keeping all others no shorter cannot decrease the value of the fairness measure. Following Definition 1, the mean, median (or any other quantile) or maximum are monotonic measures. As stated in Matl et al. (2018), if monotonicity is ensured then the optimal solution of the CVRP is composed of TSP-optimal tours. However, many interesting fairness objectives, such as the range or the variance (as it will stated in Remark 3), involve non-monotonic measures, for which route-inconsistency arises, i.e., TSP-optimality cannot be ensured, and hence the model cannot ensure that every route is of minimum length, unless these constraints are explicitly imposed in the model.

To illustrate the implications of not imposing TSP-optimality in fair vehicle routing, consider a simple VRP instance with seven nodes, six customers and a central depot, whose coordinates and demands are detailed in Table 2. Two identical vehicles are available, each with a capacity of five demand units. Table 3 reports the outcomes under a set of objective functions, including monotonic measures (mean, median, and maximum route length) and non-monotonic ones (range, Gini index, median absolute deviation ($\mathrm{MAD}_{(med)}$), and variance). For each objective, we compare solutions obtained when TSP-optimality is enforced for each route versus when it is not. Three outcomes emerge from the analysis. In the case of monotonic objectives, the optimal routing remains unchanged regardless of whether TSP-optimality is enforced, yielding identical route lengths (Tour 1: 45.03, Tour 2: 49.06, Total cost: 94.09). This solution is illustrated in Figure 1a. In contrast, for non-monotonic objectives, solutions differ substantially depending on the enforcement of TSP-optimality. When this constraint is relaxed, the solution is more flexible to minimize inequality, producing nearly equal route lengths (Tour 1: 56.57, Tour 2: 56.61, Total cost: 113.18). As depicted in Figure 1b, this results in a route inconsistency for driver two, since its tour is not TSP optimal. Conversely, when TSP-optimality is enforced, the resulting solution cannot fully equalize route lengths (Tour 1: 50.48, Tour 2: 53.52, Total cost: 104.00), but route inconsistency is eliminated as depicted in Figure 1c. This example illustrates two key trade-offs in equitable vehicle routing. First, while enforcing TSP-optimality solves the issue of route inconsistency, it does not guarantee workload consistency, since the total cost increases from the efficient baseline (94.09) to the fair solution (104.00). Second, the example shows that omitting TSP-optimality can paradoxically lead to even higher total costs (113.18), suggesting that incorporating this constraint may improve overall efficiency in certain fair-routing contexts.



(a) Monotonic measures

(b) Non-monotonic measures (no TSP-optimal)

(c) Non-monotonic measures (TSP-optimal)

Figure 1: Comparison of three routing solutions

Furthermore, our framework naturally accommodates negative entries in the weight vector $\boldsymbol{\lambda}$, a feature that is essential for representing most non-monotonic fairness measures considered in the literature. The following result shows that any non-monotonic fairness objective necessarily requires the presence of negative weights, and therefore cannot be expressed using exclusively nonnegative coefficients. This highlights a fundamental limitation of formulations restricted to nonnegative weight vectors and further motivates the generality of our approach.

Theorem 1 *Let $l_{\text{ord}}(\Phi, \boldsymbol{\lambda}) = \sum_{\ell \in K} \lambda_\ell \Phi_{(\ell)}$ be the ordered weighted objective defined over the sorted route lengths $\Phi_{(1)} \leq \cdots \leq \Phi_{(m)}$. If $l_{\text{ord}}(\Phi, \boldsymbol{\lambda})$ is non-monotonic, then there exists at least one index $j \in K$ such that $\lambda_j < 0$.*

Proof: We proceed by contraposition, proving that if $\lambda_\ell \geq 0$ for all $\ell \in K$, then l_{ord} is monotone. Let $\Phi, \Phi' \in \mathbb{R}^m_{\geq 0}$ such that $\Phi'_k = \Phi_k + \delta_k$ with $\delta_k \geq 0$ for all $k \in K$. By the monotonicity of order statistics, component-wise dominance $\Phi \leq \Phi'$ implies $\Phi_{(\ell)} \leq \Phi'_{(\ell)}$ for all $\ell = 1, \ldots, m$. The change in the measure is given by:

$$l_{\text{ord}}(\Phi', \boldsymbol{\lambda}) - l_{\text{ord}}(\Phi, \boldsymbol{\lambda}) = \sum_{\ell \in K} \lambda_\ell (\Phi'_{(\ell)} - \Phi_{(\ell)}). \tag{13}$$

Since $\lambda_\ell \geq 0$ and $(\Phi'_{(\ell)} - \Phi_{(\ell)}) \geq 0$ for all ℓ, it follows that:

$$\sum_{\ell \in K} \lambda_\ell (\Phi'_{(\ell)} - \Phi_{(\ell)}) \geq 0 \implies l_{\text{ord}}(\Phi', \boldsymbol{\lambda}) \geq l_{\text{ord}}(\Phi, \boldsymbol{\lambda}). \tag{14}$$

This confirms that the measure is monotone. Therefore, if l_{ord} is non-monotonic, there must exist at least one $\lambda_j < 0$. □

As a consequence, we have the following remark.

Remark 1 *TSP-optimality must be imposed in the presence of at least one negative weight $\lambda_j < 0$, for some $j \in K$.*

This motivates the utilization of bilevel optimization models, where, in the presence of non-monotonic measures, we must ensure that each Φ_k represents the shortest route (i.e., the optimal TSP solution) for the set of visited clients. Regarding the absolute and squared loss functions, (9) and (10) respectively, there also exist cases where we need to impose TSP-optimality in order to have optimal routes. This can be seen in Remark 2 and 3, respectively.

Remark 2 *TSP-optimality must be imposed to guarantee route efficiency when the absolute deviation loss function* (9) *is minimized.*

In order to check this, let us consider the *absolute deviation from the q-th quantile*, defined as

$$\sum_{k \in K} \left| \Phi_k - \Phi_{(q)} \right|, \tag{15}$$

where $\Phi_{(q)}$ is the q-th value in the sorted vector Φ. This metric follows the same logic as the *mean absolute deviation* but substitutes the mean with a more robust central location, the q-th quantile, see Harmati et al. (2024); Blanco et al. (2025). This measure admits a convenient $\boldsymbol{\lambda}$-vector representation as follows

$$(-1, ..., -1, \underbrace{2q - n - 1}_{q\text{-th}}, 1, ..., 1), \tag{16}$$

which have negative entries of the $\boldsymbol{\lambda}$-vector, case in which we know TSP-optimality constraints are needed to impose optimal routes.

Remark 3 *TSP-optimality must be imposed to guarantee route efficiency when the squared deviation loss function* (10) *is minimized.*

Indeed, let us consider the squared deviation with respect to a general ordered measure $l_{\text{ord}}(\Phi, \boldsymbol{\lambda})$, given by (10). As shown in Blanco et al. (2025), this expression can be reformulated as a quadratic ordered measure:

$$\mathcal{Q}(\Phi, \mathbf{A}) = \sum_{i,j \in K} a_{ij} \Phi_{(i)} \Phi_{(j)},$$

where $\mathbf{A} = (a_{ij})$ is a symmetric matrix determined by the weights $\boldsymbol{\lambda}$. For instance, in the case of the variance objective (i.e., squared deviations from the mean),

$$\sum_{i \in K} \big(\Phi_i - \frac{1}{m} \sum_{j \in K} \Phi_j\big)^2, \tag{17}$$

the matrix $\mathbf{A}$ takes the form:

$$\mathbf{A} = \begin{pmatrix} (1-\frac{1}{m}) & -\frac{1}{m} & \cdots & -\frac{1}{m} \\ -\frac{1}{m} & (1-\frac{1}{m}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{m} \\ -\frac{1}{m} & \cdots & -\frac{1}{m} & (1-\frac{1}{m}) \end{pmatrix}.$$

This matrix is the classical sample variance estimator and is known to define a non-monotonic quadratic form. This can easily be checked by taking two vectors Φ' and Φ'' such that $\Phi' \leq \Phi''$, e.g., $\Phi' = (1, 3, 5)$ and $\Phi'' = (2.5, 3, 5)$, for which $\mathcal{Q}(\Phi', \mathbf{A}) \geq \mathcal{Q}(\Phi'', \mathbf{A})$ holds. That is, increasing some components of the ordered vector Φ does not necessarily increase the value of $\mathcal{Q}(\Phi, \mathbf{A})$.

Due to this non-monotonicity, the optimization process might favor artificially constructed route lengths Φ_k that are not true shortest routes, if that allows reducing the total value of the quadratic loss. For example, increasing the length of already long routes and decreasing others might result in a lower objective, as in the previous example. Hence, without explicit bilevel constraints ensuring that each Φ_k corresponds to an optimal TSP tour, the solution may contain suboptimal routes with respect to the underlying routing problem.

The analysis of monotonicity, route inconsistency, and workload inconsistency leads to a key insight: monotonic fairness measures prevent both types of inconsistency, while non-monotonic measures can still produce workload inconsistencies even when routes are TSP-optimal. Monotonic measures, however, have limitations, e.g., the min-max objective captures less information about overall inequality than non-monotonic measures, already outlined in the introduction,. Enforcing TSP-optimality of individual routes is essential for efficient fairness. While it eliminates route inconsistency, workload inconsistency may remain, implying a “price of fairness”. Nevertheless, neglecting route optimality introduces unnecessary inefficiency, so ensuring TSP-optimality is a fundamental step toward improving workload consistency in fair vehicle routing.

4 Bilevel optimization model

This section introduces a bilevel optimization model that enforces TSP-optimality for every vehicle’s route, using a single-leader, multiple-disjoint-followers structure. The upper-level (leader) determines the allocation of customers to the k vehicles by defining a partition $\{V_1, \dots, V_m\}$ of the customer set. This assignment is modeled with binary decision variables $x_i^k \in \{0, 1\}$, where $x_i^k = 1$ indicates that customer i is assigned to vehicle k, subject to vehicle capacity constraints. Given this allocation, each lower-level follower (vehicle) must solve a TSP over its assigned customer subset. The formulation cleanly separates the strategic assignment of customers from the tactical requirement that each resulting route needs to be the shortest possible for its allocated set.

$$\min \mathcal{L}(\Phi, \nu), \tag{18a}$$

$$\text{s.t. } \nu = \sum_{k \in K} \lambda_k \Phi_k, \tag{18b}$$

$$x_0^k = 1, \qquad k \in K, \tag{18c}$$

$$\sum_{i \in V'} x_i^k \geq 1, \qquad k \in K, \tag{18d}$$

$$\sum_{k \in K} x_i^k = 1, \qquad i \in V', \tag{18e}$$

$$\sum_{i \in V} d_i x_i^k \leq Q, \qquad k \in K, \tag{18f}$$

$$\sum_{k \in K} \Phi_k \leq B, \tag{18g}$$

$$\Phi_k \leq \Phi_{k+1}, \qquad k \in K : k < m, \tag{18h}$$

$$x_i^k \in \{0, 1\}, \qquad i \in V, k \in K. \tag{18i}$$

where Φ_k represents the optimal value of the x-parametrized lower-level problem:

$$\phi_k(x^k) = \min \sum_{i \in V} \sum_{\substack{j \in V \\ j \neq i}} c_{ij} z_{ij}^k, \tag{18j}$$

$$\text{s.t. } \sum_{i \in V} z_{ij}^k = x_j^k, \qquad j \in V, \tag{18k}$$

$$\sum_{j \in V} z_{ij}^k = x_i^k, \qquad i \in V, \tag{18l}$$

$$\texttt{Subtour\_elimination}(z^k, x^k), \tag{18m}$$

$$z_{ij}^k \in \{0, 1\}. \tag{18n}$$

The objective function (18a) is designed to minimize a loss function $\mathcal{L}(\Phi, \nu)$, where the scalar quantity $\nu = \sum_{k \in K} \lambda_k \Phi_k$ represents a linearly weighted and ordered aggregation of the individual route lengths Φ_k. This formulation allows for considerable flexibility in modeling various fairness or efficiency objectives. In the particular case where the loss function $\mathcal{L}$ is the identity mapping, the objective reduces to the direct minimization of ν, and constraint (18b) becomes redundant.

Constraints (18c)–(18i) define the upper-level structure of the formulation. Constraints (18c) enforce that each driver departs from and returns to the depot, while constraints (18d) ensure that no vehicle remains unused. Constraints (18e) assign each client to exactly one driver. Let B be an upper bound on the total routing cost. Capacity limitations are imposed by constraints (18f), which guarantee that the total demand served on any route does not exceed Q. Budget constraints (18g) bound the total cost of the selected arcs by B. Finally, constraint (18h) enforces a consistent ordering of route lengths required for the objective function.

On the other hand, constraints (18j)–(18n) define the lower-level routing problem, whose purpose is to ensure that the route length Φ_k associated with each driver $k \in K$ is optimal given the client assignments determined at the upper level. This level introduces binary variables z_{ij}^k, which take value 1 if arc $(i, j) \in A$ is traversed by vehicle k, and 0 otherwise. The lower-level objective (18j) minimizes the total travel cost incurred by driver k, while constraints (18k) and (18l) enforce that vehicle k visits exactly the customers assigned to it, namely those with $x_i^k = 1$. Under the assumption that the travel costs c_{ij} satisfy the triangle inequality, and due to the cost-minimization objective, vehicle k will not visit any customer other than those with $x_i^k = 1$.

To guarantee that the resulting route forms a single Hamiltonian tour over the assigned clients and the depot, the model incorporates connectivity or subtour elimination constraints in (18m). To eliminate invalid isolated cycles disconnected from the depot, cutset inequalities that enforce global connectivity are used in our model. To this end, for every $\emptyset \neq S \subset V$ such that $0 \notin S$ and for any representative node $h \in S$ visited by vehicle k, the following constraint is added:

$$\sum_{i\in S}\sum_{j\in V\setminus S} z_{ij}^k \ \geq \ x_h^k. \tag{19}$$

This constraint ensures that each connected component containing assigned clients has at least one outgoing arc, thereby linking it to the rest of the route and ultimately to the depot. Due to the degree constraints (18k) and (18l), these cutset inequalities also guarantee that each connected component also has at least one outgoing arc.

Also, we further improve computational performance, by means of the following constraints:

$$z_{ij}^k + z_{ji}^k \ \leq \ x_i^k, \qquad \forall i,j\in V',\ k\in K, \tag{20}$$

which forbid the simultaneous selection of the arcs (i,j) and (j,i) for vehicle k unless node i is assigned to that vehicle. As a result, they reduce the search space early in the branching tree by ruling out infeasible or degenerate routing patterns that do not correspond to valid vehicle routes.

A key insight about the bilevel structure concerns the objective function's properties. If the performance measure $\mathcal{L}$ is monotonic in individual route lengths, the bilevel formulation is unnecessary, as minimizing the global objective naturally minimizes each route. The problem can then be solved as a single-level model by including constraints (18k)-(18n) and linking Φ_k to z variables:

$$\Phi_k = \sum_{i\in V}\sum_{j\in V, j\neq i} c_{ij} z_{ij}^k. \tag{21}$$

However, as demonstrated above, most fairness criteria, such as range or Gini differences, are non-monotonic. In these cases, a leader could otherwise increase a driver's route length to meet global equity goals. The bilevel formulation prevents this, ensuring drivers always follow their shortest possible routes.

An alternative model is presented that orders routes using the ℓ-sum operator, which has proven useful for handling ordered objectives (see Marín et al., 2020; Ljubić et al., 2024). This approach linearizes the sorting via the dual representation of the ℓ-sum problem, avoiding the need for the explicit ordering constraints in (18h). A formulation for our problem using the ℓ-sum approach is given below.

$$\min \mathcal{L}(\Psi,\nu), \tag{22a}$$

$$\text{s.t. } \nu = \sum_{\ell\in\Delta^+}\Delta_\ell\big((m-\ell+1)t_\ell + \sum_{k\in K} v_{k\ell}\big) + \sum_{\ell\in\Delta^-}\Delta_\ell\sum_{k\in K}\theta_{k\ell}\Psi_k, \tag{22b}$$

$$t_\ell + v_{k\ell} \geq \Psi_k, \qquad k\in K, \ell\in\Delta^+, \tag{22c}$$

$$\sum_{k\in K}\theta_{k\ell} = m-\ell+1, \qquad \ell\in\Delta^-, \tag{22d}$$

$$x_0^k = 1, \qquad k\in K, \tag{22e}$$

$$\sum_{i\in V'} x_i^k \geq 1, \qquad k\in K, \tag{22f}$$

$$\sum_{k\in K} x_i^k = 1, \qquad i\in V', \tag{22g}$$

$$\sum_{i\in V} d_i x_i^k \leq Q, \qquad k\in K, \tag{22h}$$

$$\sum_{k\in K}\Psi_k \leq B, \tag{22i}$$

$$\Psi_k = \phi_k(x^k), \qquad i\in V, k\in K, \tag{22j}$$

$$x_i^k \in \{0,1\}, \qquad i\in V, k\in K, \tag{22k}$$

$$t_\ell\in\mathbb{R}, v_{k\ell}\geq 0, \qquad k\in K, \ell\in\Delta^+, \tag{22l}$$

$$\theta_{k\ell}\in\{0,1\}, \qquad k\in K, \ell\in\Delta^-. \tag{22m}$$

where $\phi_k(x^k)$ is defined by (18k)-(18n). This is an alternative bilevel model where constraints (22j) enforce that Ψ_k is the length of the optimal TSP tour for the set of customers given by the vector x^k. Variable ν is equal to $l_{\text{ord}}(\Psi,\lambda)$ but considering the ℓ-sum representation in (22b). As observed in other studies in Location Analysis, this alternative is often more computationally robust for large-scale instances and usually provides better linear relaxation values.

5 Handling the bilevel constraints

In what follows, we first discuss how to reformulate the bilevel problems into a single-level reformulation. In problem (18), the lower-level (18j)-(18n), is not convex. Therefore, we cannot solve this bilevel problem by reformulating it using the Karush-Kuhn-Tucker (KKT) conditions nor by exploiting the strong-duality. Instead, we use the value function reformulation given as follows:

$$\min \mathcal{L}(\Phi, \nu), \tag{23a}$$

$$\text{s.t. } \nu = \sum_{k \in K} \lambda_k \Phi_k, \tag{23b}$$

$$x_0^k = 1, \qquad k \in K, \tag{23c}$$

$$\sum_{i \in V'} x_i^k \geq 1, \qquad k \in K, \tag{23d}$$

$$\sum_{k \in K} x_i^k = 1, \qquad i \in V', \tag{23e}$$

$$\sum_{i \in V} d_i x_i^k \leq Q, \qquad k \in K, \tag{23f}$$

$$\sum_{k \in K} \Phi_k \leq B, \tag{23g}$$

$$\Phi_k \leq \Phi_{k+1}, \qquad k \in K : k < m, \tag{23h}$$

$$\sum_{i \in V} z_{ij}^k = x_j^k, \qquad k \in K, j \in V, \tag{23i}$$

$$\sum_{j \in V} z_{ij}^k = x_i^k, \qquad k \in K, i \in V, \tag{23j}$$

$$\texttt{Subtour\_elimination}(z^k, x^k), \qquad k \in K, \tag{23k}$$

$$\sum_{i \in V} \sum_{j=0, j \neq i}^{n} c_{ij} z_{ij}^k = \Phi_k, \qquad k \in K, \tag{23l}$$

$$\Phi_k \leq \phi_k(x^k), \qquad k \in K, \tag{23m}$$

$$x_i^k \in \{0, 1\}, \qquad i \in V, k \in K, \tag{23n}$$

$$z_{ij}^k \in \{0, 1\}, \qquad i, j \in V : i \neq j, k \in K. \tag{23o}$$

Constraints (23m) are non-convex, hence we will try to convexify them. Note that the feasible set of the lower level in problem (18) depends on the upper-level variables x^k. In lieu of solving the lower-level problem (18j)-(18n), one can solve an alternative problem whose feasible set does not depend on the upper-level variables. Instead, this problem contains additional penalty terms in the objective function, so that it is guaranteed that the same objective value is obtained.

In what follows, we note that the lower-level problem can be formulated as a Profitable Tour Problem (PTP), a variant of the TSP introduced in Feillet et al. (2005). The PTP is an NP-hard vehicle routing problem in which the objective is to maximize net profit, defined as the total revenue collected from visited customers minus the travel costs incurred. Unlike the classical TSP, it is not required to visit all customers, instead, a solution may omit unprofitable nodes, and an optimal tour typically visits only a subset of customers that yields positive net contribution. Our PTP is solved in the same graph G, with node prizes $\tilde{\pi}_i$ and arc lengths $\tilde{c}_{ij}$ defined as

$$\tilde{\pi}_i = M \bar{x}_i^k \qquad \text{and} \qquad \tilde{c}_{ij} = \bar{x}_i^k c_{ij} + M(1 - \bar{x}_i^k) \tag{24}$$

for sufficiently large values of M, where $\bar{x}$ is a feasible leader vector. This is proven in the following proposition, where the penalization terms M^1 and M^2 are also specified.

Proposition 1 *Given the fixed values for the upper-level decision variables, $\overline{x}_i^k \in \{0, 1\}$, for all $k \in K$ and $i \in V$, satisfying constraints* (23c) - (23f), *let us define the binary decision variables $y_i^k \in \{0, 1\}$, which are equal to 1 if and only if the vehicle $k \in K$ decides to visit customer $i \in V'$. Then, for every $k \in K$, instead of solving problem*

(18l)-(18n) *one can solve the following equivalent problem,*

$$\phi_k(\bar{x}^k) = \min \sum_{i \in V} \overline{x}_i^k \sum_{j \in V, j \neq i} c_{ij} z_{ij}^k + \sum_{i \in V} M_i^1 \overline{x}_i^k (1 - y_i^k) + \sum_{i \in V} M_i^2 (1 - \overline{x}_i^k) y_i^k, \tag{25a}$$

$$\text{s.t.} \sum_{i \in V} z_{ij}^k = y_j^k, \qquad j \in V, \tag{25b}$$

$$\sum_{j \in V} z_{ij}^k = y_i^k, \qquad i \in V, \tag{25c}$$

$$\texttt{Subtour\_elimination}(z^k, y^k), \tag{25d}$$

$$z_{ij}^k \in \{0, 1\}, \qquad i, j \in V : i \neq j. \tag{25e}$$

where M_i^1 and M_i^2 are set as follows:

$$M_i^1 = \max\{c_{ri} + c_{it} - c_{rt} \ : \ r, t \in V, i \neq r, t\},$$
$$M_i^2 = \max\{c_{rt} - c_{ri} + c_{it} \ : \ r, t \in V, i \neq r, t\}.$$

Proof: To prove the desired result, it is sufficient to show that for an optimal solution y^k of the lower-level problem, $\overline{x}_i^k = y_i^k$ holds for all $i \in V$ and $k \in K$. Given that the argument applies uniformly to all vehicles, we omit explicit reference to the vehicle index $k \in K$.

Assume that $\overline{x}_i = 1$ for some $i \in V$. This means that customer i is visited in the TSP solution defined by the upper-level constraints. Let the path through i be given by $u \to i \to v$, with $u, v \in V$ and $i \neq u, v$. This covers cases where a vehicle visits only one customer ($0 \to i \to 0$) by allowing $u = 0$ and $v = 0$. If we choose $y_i = 0$, we incur a penalty cost of M_i^1. The total cost of the solution includes this penalty plus the cost of a tour that does not visit i. If we choose $y_i = 1$, the tour must include node i, adding a routing cost of $c_{ui} + c_{iv} - c_{uv}$. Since $M_i^1 = \max_{r,t \in V, i \neq r,t} \{c_{ri} + c_{it} - c_{rt}\}$, and our specific nodes u, v satisfy the condition $i \neq u, v$, then $M_i^1 \geq c_{ui} + c_{iv} - c_{uv}$. Therefore, paying the penalty M_i^1 is always at least as expensive than including node i in the tour. Thus, there exist an optimal solution such that $y_i = 1$.

Now, consider the case where a node $s \in V$ is not prescribed to be visited by the upper level, meaning $\overline{x}_s = 0$. We must show that any optimal solution will have $y_s = 0$. Assume by contradiction that an optimal solution exists where $y_s = 1$. This implies that the tour created by the z_{ij}^k variables passes through node s. Let this path be $\dots \to u \to s \to v \to \dots$, where u is a prescribed node ($\overline{x}_u = 1$). The decision to visit s incurs a penalty M_s^2 in the objective function. Crucially, in (25a), the cost c_{us} is included (as $\overline{x}_u = 1$), but the cost c_{sv} is not (as $\overline{x}_s = 0$).

To ensure that visiting s is never optimal, the cost of the detour ($c_{us} + M_s^2$) must be greater than to the cost of the direct arc (u, v) that the leader expects. By our definition $M_s^2 = \max_{r,t \in V, s \neq r,t}\{c_{rt} - c_{rs} + c_{st}\}$, and since $s \neq u$ and $s \neq v$ in a valid TSP tour, the pair $(r = u, t = v)$ is a candidate for the maximum. Thus:

$$M_s^2 \geq c_{uv} - c_{us} + c_{sv}$$

Substituting this into the follower's perceived cost for the path through s:

$$c_{us} + M_s^2 \geq c_{us} + (c_{uv} - c_{us} + c_{sv}) = c_{uv} + c_{sv}$$

Since $c_{sv} \geq 0$, it follows that $c_{uv} + c_{sv} \geq c_{uv}$. This proves that the penalized cost of the detour is at least as expensive as the original arc (u, v), and thus an optimal solution of problem (25) exists in which $y_s = 0$.

Furthermore, if the follower visits a sequence of unassigned nodes, each node s_g independently triggers a penalty $M_{s_g}^2$. Since each penalty is large enough to offset the "free" arc used to exit that node, the cumulative penalty will always outweigh any distance savings. Thus, any such deviation would result in a strictly more expensive solution, contradicting optimality. It must be that we can always find a solution such that $y_s = 0$ whenever $\overline{x}_s = 0$. □

As a consequence of Proposition 1, since the feasible region of problem (25) does not depend on x variables we can represent the function ϕ_k as follows in Corollary 1 in terms of routes.

Corollary 1 *Let $\mathcal{T}$ be the set of all the routes in graph G, we denote by $T \in \mathcal{T}$ an arbitrary route, with V_T and A_T being the set of nodes and arcs visited and traversed by the route, respectively. Then, for every $k \in K$, the value function ϕ_k in problem* (18) *is equivalent to*

$$\phi_k(x^k) = \min_{T=(V_T, A_T)} \sum_{(i,j) \in A_T} x_i^k c_{ij} + \sum_{i \notin V_T} M_i^1 x_i^k + \sum_{i \in V_T} M_i^2 (1 - x_i^k). \tag{26}$$

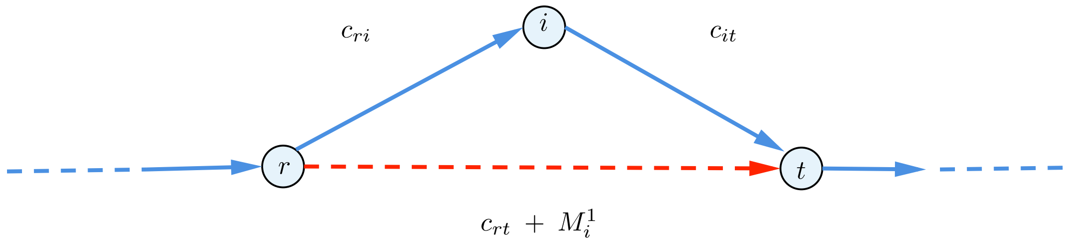


Figure 2: Subset penalization scenario

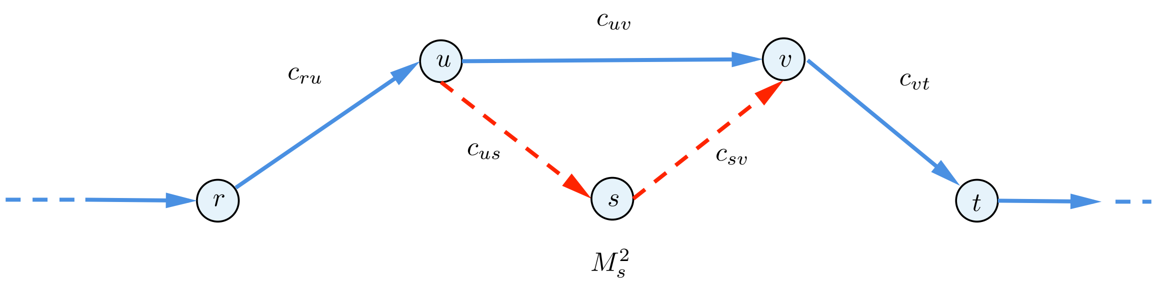


Figure 3: Superset penalization scenario

That way, we convexified the value function $\phi_k(x^k)$ by representing it as a minimum of affine functions in x^k. By replacing constraint (23m) in the model (23), with the following ones:

$$\Phi_k \leq \sum_{(i,j)\in A_T} x_i^k c_{ij} + \sum_{i\notin V_T} M_i^1 x_i^k + \sum_{i\in V_T} M_i^2(1 - x_i^k), \quad (V_T, A_T) \in \mathcal{T}, \tag{27}$$

we obtain a single-level reformulation for the first model. Since for every $i \in T$, at most one outgoing arc belongs to A_T, constraints (23m) can be rearranged to

$$\Phi_k \leq \sum_{i\in V_T} \sum_{(i,j)\in A_T} \left(x_i^k c_{ij} + M_i^2(1 - x_i^k)\right) + \sum_{i\notin V_T} M_i^1 x_i^k, \quad (V_T, A_T) \in \mathcal{T}. \tag{28}$$

It is important to note that, for the theoretical validity of our model, the cuts defined in (27) are required only for those components of the $\boldsymbol{\lambda}$-vector that are negative. However, in the case of criteria such as the general range, which includes zero-valued components in the $\boldsymbol{\lambda}$-vector, there may exist routes that are not explicitly represented in the objective function, particularly those lying between the minimum and maximum values. Consequently, these routes are not necessarily required to be optimal. To address this potential non-optimality, one can introduce the aforementioned cuts for all vehicles across all criteria, including those associated with zero-valued components of $\boldsymbol{\lambda}$. Although this extension is not strictly necessary from a modeling perspective, our computational experiments reveal that incorporating these additional cuts significantly strengthens the model's bounds. Therefore, we impose these optimality-enforcing cuts for all entries of the $\boldsymbol{\lambda}$-vector, not solely for the negative ones.

For the ℓ-sum model, the value function reformulation follows an identical logic, yielding the bilevel cuts of the form:

$$\Psi_k \leq \sum_{(i,j)\in A_T} x_i^k c_{ij} + \sum_{i\notin V_T} M_i^1 x_i^k + \sum_{i\in V_T} M_i^2(1 - x_i^k), \quad (V_T, A_T) \in \mathcal{T}. \tag{29}$$

Separation of the Bilevel Cuts

Since the number of constraints (27) is exponential, we solve formulation (23) using a dynamic separation procedure. Given a current master solution $(\bar{x}^k, \bar{\Phi}_k)$, we solve the following TSP subproblem to find the optimal lower-level tour for the selected set of customers:

$$\phi_k(\bar{x}^k) = \min \sum_{i\in V} \sum_{j\in V, j\neq i} c_{ij} z_{ij}^k \tag{30a}$$

$$\text{s.t.} \sum_{i\in V} z_{ij}^k = \bar{x}_j^k, \qquad j \in V, \tag{30b}$$

$$\sum_{j \in V} z_{ij}^k = \bar{x}_i^k, \qquad i \in V, \tag{30c}$$

$$\texttt{Subtour\_elimination}(z^k, x^k), \tag{30d}$$

$$z_{ij}^k \in \{0,1\}, \qquad i,j \in V, i \neq j. \tag{30e}$$

Let $(\bar{V}_T, \bar{A}_T)$ represent the optimal tour obtained from (30). If the optimal tour length $TSP(\bar{V}_T) < \bar{\Phi}_k$, the corresponding bilevel cut is violated. We then add the valid inequality to the master problem in the form of (27), associated to $(\bar{V}_T, \bar{A}_T)$.

The penalization terms M^1 and M^2 can be significantly tightened by calculating them relative to the specific tour T identified during separation. Thus, the cuts in (27) can be also implemented as:

$$\Phi_k \leq \sum_{(i,j) \in A_T} x_i^k c_{ij} + \sum_{i \notin V_T} M_i^{1,T} x_i^k + \sum_{i \in V_T} M_i^{2,T}(1 - x_i^k), \quad (V_T, A_T) \in \mathcal{T}, \tag{31}$$

where the tour-dependent penalties are defined as:

$$M_i^{1,T} = \max\{c_{ri} + c_{it} - c_{rt} : r,t \in V_T, i \neq r,t\},$$
$$M_i^{2,T} = \max\{c_{rt} - c_{ri} - c_{it} : r,t \in V_T, i \neq r,t\}.$$

Therefore, we consider two distinct cutting-plane strategies: a `GLOBAL` approach, in which penalties are precomputed, and a `LOCAL` approach, where penalties are dynamically tightened within a callback based on the current tour. It is important to clarify that the distinction between `GLOBAL` and `LOCAL` refers specifically to how the penalties are computed, not to the validity of the resulting cuts. Although cuts generated within callbacks are often termed local in the literature when they are only valid within the subtree where they are generated, the cuts produced by our `LOCAL` approach are globally valid. Nonetheless, because they are generated on the fly during the branch-and-bound process and depend on the incumbent tour at a given node, in which our penalties are adjusted, we refer to this strategy as `LOCAL`.

For the ℓ-sum model, the separation procedure follows the same rationale. In this case, bilevel cuts under the `GLOBAL` strategy are given by (29), whereas those corresponding to the `LOCAL` strategy take the form:

$$\Psi_k \leq \sum_{(i,j) \in A_T} x_i^k c_{ij} + \sum_{i \notin V_T} M_i^{1,T} x_i^k + \sum_{i \in V_T} M_i^{2,T}(1 - x_i^k), \quad (V_T, A_T) \in \mathcal{T}. \tag{32}$$

6 Improvements

This section details several computational enhancements aimed at mitigating the complexity inherent in the previously defined bilevel models. The presence of three-index binary variables, which scale with the number of arcs and vehicles, often becomes a bottleneck for solver performance. We first propose a projection technique to derive sparser formulations by substituting these high-dimensional variables with specific families of valid inequalities, such as integer L-shaped and logic-based Benders cuts. Subsequently, we establish a methodology for deriving robust lower and upper bounds for route lengths. These bounds are instrumental in pruning the search space and accelerating the convergence of the branch-and-cut algorithm, particularly for larger instances where tractability is a primary concern.

Deriving Sparser Models

The value function reformulation (23) necessitates the inclusion of all lower-level constraints, including the z_{ij}^k binary variables which scale at $O(n^2 m)$, potentially hindering model tractability. However, these routing variables can be projected out of the formulation and replaced by specific constraints that characterize the TSP cost as a function of the assignment variables x.

In this section, we show that in both of the proposed models (simple ordering and ℓ-sum), these z variables can be projected out and replaced by certain constraints, resulting in two much sparser formulations. In both models, degree constraints, subtour elimination, and integrality constraints are removed. These are constraints from (23i) to (23l), which can be replaced in model (23) with the following integer L-shaped cuts (see, e.g., Laporte and Louveaux, 1998):

$$\Phi_k \geq \underline{\Phi}_k + (\mathsf{TSP}(V_T) - \underline{\Phi}_k) \left(\sum_{i \in V_T} x_i^k - |V_T| + 1 \right), \quad (V_T, A_T) \in \mathcal{T}, k \in K, \tag{33}$$

where $\underline{\Phi}_k$ denotes a global lower bound on the k-th route length. These cuts serve as "no-good" constraints, this is, if vehicle k is assigned exactly the set of nodes V_T, the lower bound on Φ_k is forced to the optimal TSP value $\mathsf{TSP}(V_T)$. If even one node from V_T is excluded, the constraint reverts to the global lower bound, ensuring validity across all partitions.

A similar logic applies to the ℓ-sum model (22), where they are replaced by

$$\Psi_k \geq \underline{\Psi} + (\mathsf{TSP}(V_T) - \underline{\Psi})(\sum_{i \in V_T} x_i^k - |V_T| + 1), \quad (V_T, A_T) \in \mathcal{T}, k \in K, \tag{34}$$

with $\underline{\Psi}$ representing a global lower bound on *any route* in an optimal solution. How to obtain the values of $\underline{\Phi}_k$ and $\underline{\Psi}$ is discussed in the next section.

Alternatively, a family of logic-based TSP optimality cuts can be utilized to strengthen the formulation (see, e.g., in Fachini and Armentano, 2020). These cuts rely on the triangle inequality to provide a valid lower bound even when the current customer set is a subset of V_T. For model (23), the cuts are defined as:

$$\Phi_k \geq \mathsf{TSP}(V_T) - \left(\sum_{i \in V_T} (1 - x_i^k)(\max_{ij \in \delta^+(i)} \{c_{ij}\} + \max_{ji \in \delta^-(i)} \{c_{ji}\}) \right), \quad (V_T, A_T) \in \mathcal{T}, k \in K. \tag{35}$$

The term being subtracted accounts for the maximum possible reduction in tour length if node i is removed from the set V_T. By subtracting the two most expensive incident arcs for each excluded node, the inequality remains valid while providing a tighter bound than the L-shaped alternative. Similarly, for the ℓ-sum model (22), the cuts enforce the lower bound on Ψ_k variables as follows:

$$\Psi_k \geq \mathsf{TSP}(V_T) - \left(\sum_{i \in V_T} (1 - x_i^k)(\max_{ij \in \delta^+(i)} \{c_{ij}\} + \max_{ji \in \delta^-(i)} \{c_{ji}\}) \right), \quad (V_T, A_T) \in \mathcal{T}, k \in K. \tag{36}$$

We note that the bilevel cuts are already expressed in terms of the Φ_k (respectively, Ψ_k) variables. Consequently, by combining (27) (respectively, (29)) with the lower-bound cuts introduced in this section, we obtain valid and parsimonious formulations.

Deriving bounds for the route lengths

The performance of the projected formulations is strongly influenced by the numerical tightness of the admissible region defined for the route-length variables. Deriving meaningful lower and upper bounds on route costs is therefore essential, as these bounds limit the feasible values of Φ_k and Ψ_k and ensure consistency between the routing subproblem and the projected representation. Carefully calibrated bounds enhance the stability of the formulation, reinforce the effectiveness of the associated valid inequalities, and substantially improve the solver's convergence behavior.

Lower bounds for Φ_k are derived through a two-step relaxation process. The first step involves constructing a relaxation model to estimate how many customers can be feasibly served by $m-1$ vehicles, that is, all vehicles but one. For each node $i \in V$, we introduce a binary variable α_i which is equal to 1 if the node is selected for service, and 0 otherwise. Additionally, for each node i and vehicle $j \in \{1, \ldots, K-1\}$, a binary variable β_{ij} is introduced, indicating whether node i is assigned to vehicle j. The model maximizes the total number of selected nodes as follows,

$$\max \sum_{i \in V'} \alpha_i, \tag{37}$$

$$\text{s.t.} \sum_{j=1}^{m-1} \beta_{ij} = \alpha_i, \qquad i \in V', \tag{38}$$

$$\sum_{i \in V'} d_i \beta_{ij} \leq Q, \qquad j \in \{1, \ldots, m-1\}. \tag{39}$$

Let obj^* be the optimal value of the objective function. We then define the number of unserved nodes as $a = \max\{1, n - obj^*\}$. With a determined, we construct a second model to compute a lower bound on the tour cost needed to visit exactly a nodes. This is formulated as a Prize Collecting TSP over the set of all customers, where we minimize the tour length and impose two constraints: the number of served customers must be a and the capacity of served customers should not exceed Q including cardinality constraint. If this model is solved successfully, the optimal tour length serves as a conservative lower bound for the route lengths, i.e., Φ_k for all $k \in K$. In case the second model fails to find an optimal solution due to a time limit, a second bound is computed as $c_{0i} + c_{i0}$, which corresponds to the shortest distance from the depot to any other node i. The final output is the maximum of the two.

To derive valid upper bounds for the ordered cost components Φ_k, we proceed as follows. A natural upper bound on individual route lengths is given by the optimal value of the classical TSP that ignores demand constraints. However, solving such a problem is itself NP-hard, and thus may not always be computationally feasible. When solvable within a prescribed time limit, this value is adopted as a valid upper bound. Otherwise, we fall back on a greedy heuristic to construct a feasible TSP tour, which yields a looser but efficient approximation. Additionally, we propose a more refined procedure to compute an upper bound. We initialize the bound UB to 0. Then, for each client node $i \in V'$, we define the subset $V_i = V' \setminus \{i\}$ and solve a TSP over V_i, denoted by $\mathsf{TSP}(V_i)$. The candidate upper bound is updated as: $UB = \max\{UB,\ c_{0i} + c_{i0},\ \mathsf{TSP}(V_i)\}$. The final upper bound on route lengths is then given by the minimum between this value and the TSP solution over the full client set (if available).

7 Computational study

This section presents a comprehensive computational evaluation of the proposed mathematical formulations. The experimental setup is not intended to compare alternative fairness measures or to evaluate the impact of enforcing TSP-optimality of the individual routes within CVRP-based models, as these issues have already been thoroughly discussed in the literature (see, e.g., Matl et al., 2018; Halvorsen-Weare and Savelsbergh, 2016). Instead, the purpose of these experiments is to assess and compare the computational performance and solution quality of the proposed formulations under a common and controlled benchmark.

All models were implemented in Python (version 3.10) using the Gurobi Optimizer (version 9.5) as the core Mixed-Integer Programming (MIP) solver. Computational experiments were executed on a MacPro server equipped with a 2.7 GHz Intel Xeon W processor (24 cores) and 192 GB of RAM. In the experiments 8 cores were used. To maintain practical relevance and ensure a fair comparison, a maximum runtime of 3,600 seconds (1 hour) was imposed on each instance.

Experimental design

To the best of our knowledge, the work by van Rossum et al. (2025) is the only existing study that enforces TSP-optimality in a provably optimal manner for the Range VRP. Consequently, we adopt their benchmark instances for comparative purposes, which are publicly available at `https://github.com/BartvanRossum/EfficientBranchingRules`. These instances were constructed following the methodology of Matl et al. (2018), derived from the X-n641-k35 instance of CVRPlib (see Uchoa et al., 2017).

For each instance size $|V| \in \{15, 20, 25\}$, a set of 20 instances was built by selecting $|V| + 1$ customer locations from the original set, with the first location designated as the depot. Each instance considers 5 vehicles with capacity Q calibrated such that all vehicles are required to satisfy the total demand. The budget B is set to 110% of the cost of the standard CVRP cost-efficient solution. All models were warm-started using the optimal solution of the classic cost-minimizing CVRP. We evaluate four formulations:

- Φ dense: Ordered formulation (18) with the routing constraints (23i) to (23l).
- Φ sparse: Ordered formulation (18) with cuts (33) and (35) instead of the routing constraints.
- Ψ dense: ℓ-sum formulation (22) with the routing constraints (23i) to (23l).
- Ψ sparse: ℓ-sum formulation (22) with cuts (34) and (36) instead of the routing constraints.

The computational framework compares two cutting-plane strategies for penalty constraints: a static `GLOBAL` approach (where the cut coeffcients are precomputed before starting branching) and a dynamic `LOCAL` approach (where the calculation of cut coefficients is embedded within callback functions). This yields eight final configurations for testing. For dense formulations, cutset inequalities enforce subtour elimination. In the sparse formulations, a combination of L-shaped and logic-based Benders cuts, namely (33) and (35) for the Φ models, and (34) and (36) for the Ψ models, was applied. All models are MIP-started with an initial solution coming from the optimal cost-efficient CVRP solution.

Seven statistical measures, covering both monotonic and non-monotonic objectives, are used as optimization criteria to capture classical efficiency considerations as well as fairness-oriented workload balancing. While the study primarily targets non-monotonic fairness objectives, monotonic measures are also included to have a full perspective of the performance of the proposed formulations. The monotonic objectives considered are `MAX`, which minimizes the maximum route cost, and `MEDIAN`, which minimizes the median route cost. The non-monotonic fairness measures include `RANGE` (the difference between the maximum and minimum route costs), `GINI` (the Gini coefficient of inequality), MADMED (mean absolute deviation from the median or $\text{MAD}_{(median)}$), MADMIN (mean absolute deviation from the minimum or $\text{MAD}_{(min)}$), and VARIANCE (the variance of route costs). Formal definitions of these measures are provided in Table 1, except for the variance which is defined in (17). Four performance metrics, are recorded:

- RUNTIME: Computational time in seconds averaged among all 20 instances per fixed $|V|$.
- $\#_{\texttt{opt}}$: Total number of instances in which optimality was certified.
- $\texttt{GAP}_{\texttt{opt}}$: Relative optimality gap between best upper and lower bounds achieved by the model averaged among all 20 instances.
- $\texttt{GAP}_{\texttt{upper-best}}$: Relative gap averaged among all 20 instances to the best known primal bound across all models, defined as
$$\texttt{GAP}_{\texttt{upper-best}} = \frac{f_\sigma - f_*}{f_*} \times 100,$$
where f_σ is the objective value obtained by formulation σ and f_* is the best primal upper bound found for that instance and metric.

The latter metric is included because, for larger instances, the non-negativity of all measures often allows good-quality upper-bound solutions to be found while the corresponding lower bounds remain equal to zero. In such cases, a zero value of $\texttt{GAP}_{\texttt{upper-best}}$ indicates that the formulation attains the best solution identified across all configurations. Therefore, similar to $\texttt{GAP}_{\texttt{opt}}$, the closer to zero the better the model performance. Aggregate results for all instances under these performance metrics are reported in Table 4, where the smallest value per row has been highlighted.

Computational results

First, we begin by evaluating the performance of the two penalty cut generation strategies, GLOBAL and LOCAL, to determine the most efficient algorithmic approach. Table 5 presents the average number of bilevel cuts introduced by each strategy, categorized by instance size ($|V|$), formulation type (Φ vs. Ψ), and sparsity structure (DENSE vs. SPARSE).

SIZE	Φ MODELS		Ψ MODELS	
	DENSE	SPARSE	DENSE	SPARSE
	GLOBAL / LOCAL	GLOBAL / LOCAL	GLOBAL / LOCAL	GLOBAL / LOCAL
15	31.01 / 24.19	223.23 / 144.20	53.06 / 40.96	203.53 / 146.03
20	199.74 / 106.39	1074.36 / 614.63	261.39 / 142.99	846.78 / 528.99
25	372.23 / 155.16	1437.43 / 684.69	593.83 / 273.91	950.08 / 453.91

Table 5: Average number of bilevel cuts introduced by GLOBAL and LOCAL penalty strategies across different model configurations and instance sizes.

The results reveal a consistent and striking pattern: the LOCAL strategy introduces substantially fewer cuts, typically around half, compared to the GLOBAL approach. This reduction is particularly pronounced for sparse formulations and larger instances. For example, in the sparse Φ models at $|V| = 25$, the average number of cuts drops drastically from 1437.43 to 684.69. By dynamically adjusting the penalization terms inside the callback, the LOCAL strategy allows for more refined cuts. Consequently, the remainder of our analysis focuses exclusively on the LOCAL configuration. Still, the interested reader might find in Table 4 all the information summarizing the aggregated computational results both strategy, reporting runtimes, optimality gaps ($\texttt{GAP}_{\texttt{opt}}$), and gaps to the best-known upper bound ($\texttt{GAP}_{\texttt{upper-best}}$). The data demonstrates a clear size-dependent hierarchy regarding formulation effectiveness.

From Table 4 we observe that for the smallest instances ($|V| = 15$), all models terminate certifying optimality for most cases within the time limit. Sparse Φ models consistently achieve the fastest solution times (often below 30 seconds) and zero optimality gaps across all objectives, solving instances up to 100 times faster than their dense or Ψ-based counterparts. It is worth noting that prior studies, such as those by Halvorsen-Weare and Savelsbergh (2016) and Matl et al. (2018), are based on enumeration methods and are limited to instances with at most 10 nodes. This underscores the inherent difficulty of embedding fairness considerations optimally into a canonical vehicle routing problem (VRP), especially when non-monotonic measures are involved.

As problem size increases to $|V| = 20$, the computational burden grows substantially. Sparse Φ formulations continue to demonstrate superior efficiency, maintaining near-zero optimality gaps and requiring 600 to 1,200 seconds on average, compared to 1,800 to 2,500 seconds for competing approaches. However, at the largest scale ($|V| = 25$), a fundamental shift occurs. The problem becomes substantially harder, and nearly all models reach the 3,600-second time limit without certifying optimality. Although optimal certification is rarely achieved at $|V| = 25$, dense Ψ models achieve

	MEASURE	SIZE	Φ MODELS DENSE GLOBAL	Φ MODELS DENSE LOCAL	Φ MODELS SPARSE GLOBAL	Φ MODELS SPARSE LOCAL	Ψ MODELS DENSE GLOBAL	Ψ MODELS DENSE LOCAL	Ψ MODELS SPARSE GLOBAL	Ψ MODELS SPARSE LOCAL
RUNTIME ($\#_{opt}$)	MAX	15	1197.61 (20)	1377.13 (18)	**10.85** (20)	190.29 (19)	1635.04 (20)	1694.34 (19)	126.88 (20)	133.98 (20)
		20	1948.81 (13)	2056.85 (11)	**663.75** (20)	893.16 (20)	1948.64 (14)	2274.32 (12)	1888.67 (17)	1893.82 (17)
		25	3602.99 (0)	3602.57 (0)	3961.11 (0)	3617.60 (0)	3604.23 (0)	3603.90 (0)	3614.39 (0)	3621.21 (0)
	MEDIAN	15	116.26 (20)	93.00 (20)	27.38 (20)	**26.96** (20)	610.85 (19)	494.63 (20)	358.94 (20)	278.06 (20)
		20	792.79 (18)	862.56 (19)	**741.28** (20)	1086.46 (19)	1417.18 (15)	1306.10 (16)	3533.92 (1)	3529.18 (2)
		25	**3423.38** (2)	3533.83 (1)	3625.12 (0)	3624.07 (0)	3604.89 (0)	3603.92 (0)	3614.03 (0)	3612.81 (0)
	RANGE	15	497.08 (20)	571.59 (20)	17.97 (20)	**17.36** (20)	2027.39 (15)	2124.45 (15)	121.99 (20)	115.20 (20)
		20	1877.88 (13)	1842.01 (14)	**572.83** (20)	690.46 (20)	2459.52 (10)	2484.00 (12)	2207.46 (14)	2419.09 (11)
		25	3608.03 (0)	3604.76 (0)	3632.76 (0)	3617.95 (0)	3603.94 (0)	3604.50 (0)	3617.30 (0)	3617.20 (0)
	GINI	15	413.31 (20)	349.09 (20)	**18.88** (20)	18.93 (20)	1638.36 (17)	1868.92 (14)	119.96 (20)	121.17 (20)
		20	1898.29 (13)	1845.69 (13)	**654.31** (20)	837.07 (19)	2382.21 (12)	2253.63 (11)	2489.65 (11)	2653.00 (10)
		25	3603.36 (0)	3605.72 (0)	3624.34 (0)	3618.79 (0)	3604.73 (0)	3603.58 (0)	3620.60 (0)	3619.35 (0)
	MADMED	15	381.28 (20)	400.36 (20)	**19.80** (20)	20.88 (20)	2420.27 (12)	2585.83 (11)	241.46 (20)	250.65 (20)
		20	1889.05 (14)	1825.57 (14)	**817.57** (20)	1083.52 (19)	2653.76 (9)	2761.37 (9)	3302.74 (5)	3253.30 (6)
		25	3605.44 (0)	3606.96 (0)	3618.77 (0)	3624.85 (0)	3604.03 (0)	3603.66 (0)	3618.90 (0)	3615.25 (0)
	MADMIN	15	457.94 (20)	462.73 (20)	20.89 (20)	**20.42** (20)	1616.30 (17)	1692.50 (17)	144.38 (20)	165.34 (20)
		20	1846.16 (14)	1890.94 (14)	**994.03** (19)	1070.73 (20)	2308.92 (12)	2282.09 (11)	2746.04 (9)	2685.71 (10)
		25	3602.60 (0)	3605.87 (0)	3617.89 (0)	3620.34 (0)	3603.40 (0)	3604.33 (0)	3621.92 (0)	3622.22 (0)
	VARIANCE	15	2029.22 (14)	2001.59 (18)	27.69 (20)	**25.22** (20)	3240.40 (4)	3070.06 (7)	130.33 (20)	125.37 (20)
		20	2524.02 (11)	2453.35 (11)	1254.03 (19)	**1138.88** (18)	3066.05 (7)	2659.11 (8)	2460.02 (13)	2588.10 (10)
		25	3603.14 (0)	3602.36 (0)	3612.30 (0)	3613.99 (0)	3604.96 (0)	3604.16 (0)	3615.83 (0)	3617.66 (0)
GAP_{opt}	MAX	15	**0.00**	0.53	**0.00**	0.08	**0.00**	0.13	**0.00**	**0.00**
		20	7.67	7.21	**0.00**	**0.00**	4.71	5.11	0.25	0.15
		25	37.34	38.86	90.55	85.95	32.33	**31.13**	84.67	90.04
	MEDIAN	15	**0.00**	**0.00**	**0.00**	**0.00**	0.40	**0.00**	**0.00**	**0.00**
		20	1.61	0.30	**0.00**	0.19	4.18	3.81	47.98	52.29
		25	**28.45**	31.85	87.64	88.17	35.71	35.16	92.40	92.50
	RANGE	15	**0.00**	**0.00**	**0.00**	**0.00**	7.52	6.95	**0.00**	**0.00**
		20	27.79	21.88	**0.00**	**0.00**	31.89	34.08	13.86	18.94
		25	98.95	99.40	100.00	100.00	**95.25**	96.66	100.00	100.00
	GINI	15	**0.00**	**0.00**	**0.00**	**0.00**	4.11	5.72	**0.00**	**0.00**
		20	27.89	29.54	**0.00**	5.00	33.95	36.01	16.22	24.20
		25	96.78	97.23	100.00	100.00	**95.91**	96.78	100.00	100.00
	MADMED	15	**0.00**	**0.00**	**0.00**	**0.00**	16.62	21.83	**0.00**	**0.00**
		20	27.37	27.39	**0.00**	5.00	43.24	43.21	47.66	41.04
		25	98.29	**97.64**	100.00	100.00	97.70	98.60	100.00	100.00
	MADMIN	15	**0.00**	**0.00**	**0.00**	**0.00**	6.96	7.83	**0.00**	**0.00**
		20	23.02	23.54	**5.00**	**5.00**	37.58	40.57	23.58	19.09
		25	96.77	99.40	100.00	100.00	**94.76**	96.57	100.00	100.00
	VARIANCE	15	18.44	8.51	**0.00**	**0.00**	52.15	44.99	**0.00**	**0.00**
		20	45.00	38.42	**5.00**	9.61	48.61	44.34	15.04	19.54
		25	99.23	100.00	100.00	100.00	**99.28**	**98.31**	100.00	100.00
$GAP_{upper\text{-}best}$	MAX	15	**0.00**	0.16	**0.00**	0.08	**0.00**	**0.00**	**0.00**	**0.00**
		20	0.73	0.51	**0.00**	**0.00**	**0.00**	0.01	**0.00**	**0.00**
		25	5.35	5.66	7.49	7.63	1.51	**0.27**	7.20	6.41
	MEDIAN	15	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	0.51	**0.00**	**0.00**	0.03	**0.00**	**0.00**	0.05	1.14
		25	3.47	6.93	13.82	15.08	1.63	**1.48**	11.13	12.43
	RANGE	15	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	0.93	3.16	**0.00**	0.20	**0.00**	**0.00**	**0.00**	**0.00**
		25	18.43	21.77	35.95	38.26	**4.22**	5.79	40.99	37.89
	GINI	15	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	3.06	3.44	**0.00**	1.89	**0.00**	0.07	**0.00**	0.47
		25	17.80	17.02	34.00	36.39	**4.44**	11.49	37.59	38.26
	MADMED	15	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	3.55	3.35	**0.00**	1.26	0.90	1.74	**0.00**	1.75
		25	15.89	13.03	32.18	35.67	**8.91**	10.75	39.06	33.64
	MADMIN	15	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	3.16	2.59	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.60
		25	22.11	17.68	53.11	52.06	**2.53**	11.40	50.56	46.44
	VARIANCE	15	1.69	0.33	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		20	5.90	4.75	2.58	3.90	0.02	0.01	**0.00**	1.40
		25	25.46	27.21	55.84	55.15	**5.11**	9.27	55.50	49.43

Table 4: Aggregated computational results for model comparisons. Reported values correspond to totals and averages over 20 instances for each problem size and objective. Runtimes are measured in seconds, and optimality gaps are expressed in percentage terms. The number of instances solved to optimality is reported in parentheses. Best-performing results are highlighted in bold.

significantly smaller $\texttt{GAP}_{\texttt{opt}}$ (30–35%) and drastically reduce the $\texttt{GAP}_{\texttt{upper-best}}$. This highlights a critical trade-off: while the Φ structure facilitates efficient cutting-plane strengthening at moderate scales, the Ψ formulation is decisive for high-quality heuristic performance on larger instances.

As a general comment, the results confirm that non-monotonic fairness objectives are substantially more challenging than monotonic ones. For $|V| = 25$ (the most difficult instances), optimality gaps for MAX and MEDIAN remain below 40% for the best formulations, while gaps for fairness measures like RANGE, GINI, and VARIANCE often exceed 95%. This pattern holds across all formulations and cut strategies, highlighting the intrinsic difficulty of balancing equity against efficiency in such cases. Among the fairness-oriented objectives, VARIANCE emerges as particularly challenging, consistently producing the largest optimality gaps and the longest runtimes. Measures such as RANGE and GINI, while still computationally demanding, exhibit comparatively better performance in terms of both solution quality and computational time. However, as it will be discussed in the comments, the final choice of formulation is driven primarily by instance size rather than by the specific fairness measure considered.

To provide a comprehensive, instance-level view of the solver's performance under the LOCAL strategy, we analyze the Empirical Cumulative Distribution Functions (ECDFs) for runtimes and gaps.

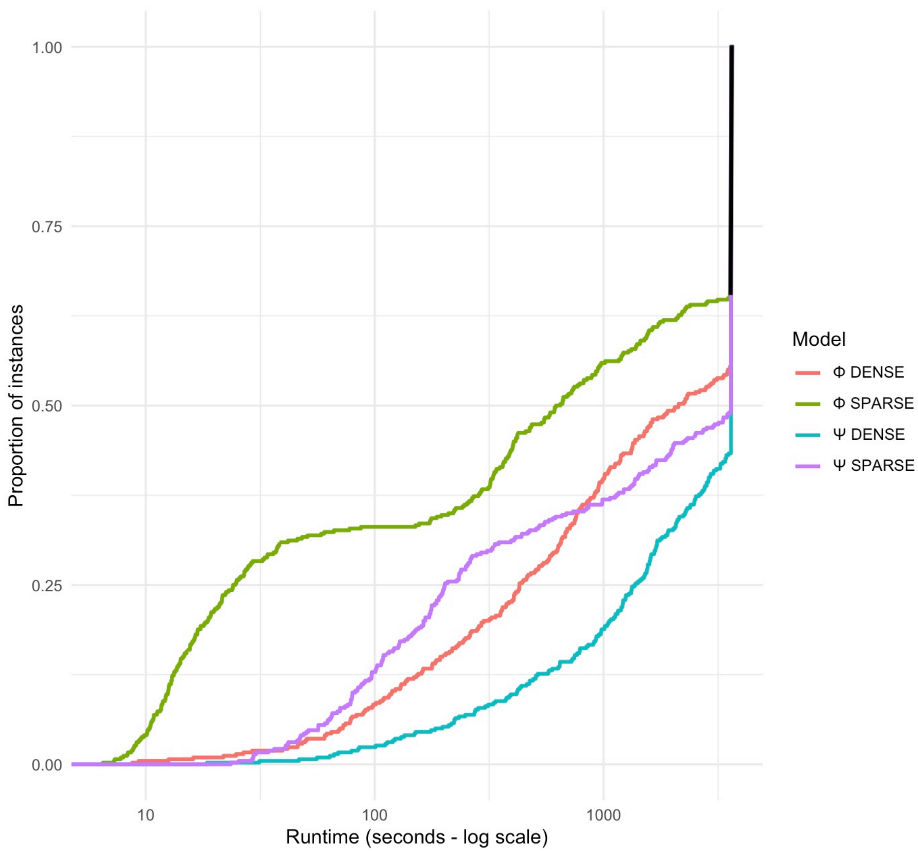


Figure 4: ECDF of runtimes (seconds, log scale) across all instances and objectives under the LOCAL strategy. Higher and leftward curves indicate faster solving.

Figure 4 presents the ECDF of runtimes. The sparse Φ model (green) exhibits a pronounced early rise, reaching approximately 30% of instances solved in under 10 seconds and 65% before the time limit. The sharp vertical jump at 3,600 seconds across all curves visually represents the heavy concentration of timeouts, confirming the ceiling effect observed for $|V| = 25$ instances.

Figure 5 shows the ECDF of optimality gaps. Both sparse Φ (green) and sparse Ψ (purple) models achieve a 0% gap in a high proportion of instances (up to 65% and 50%, respectively). However, the sparse Φ curve flattens significantly between 0% and 95%, indicating a bimodal behavior: instances are either solved to optimality or time out with massive gaps. Conversely, the dense Ψ model (cyan) exhibits a more uniform degradation and maintains a favorable profile at intermediate gap values (25–75%).

Finally, Figure 6 evaluates primal solution quality via the $\mathtt{GAP}_{\mathtt{upper\text{-}best}}$ metric. Here, the dense Ψ formulation (cyan) strongly dominates. Its curve rises sharply near zero, with over 85% of instances achieving a gap of less than 5% to the best-known solution. This reinforces our aggregated findings: sparse Φ is highly efficient for optimality certification in tractable instances, but dense Ψ delivers consistently higher-quality feasible routing plans when computational limits are reached.

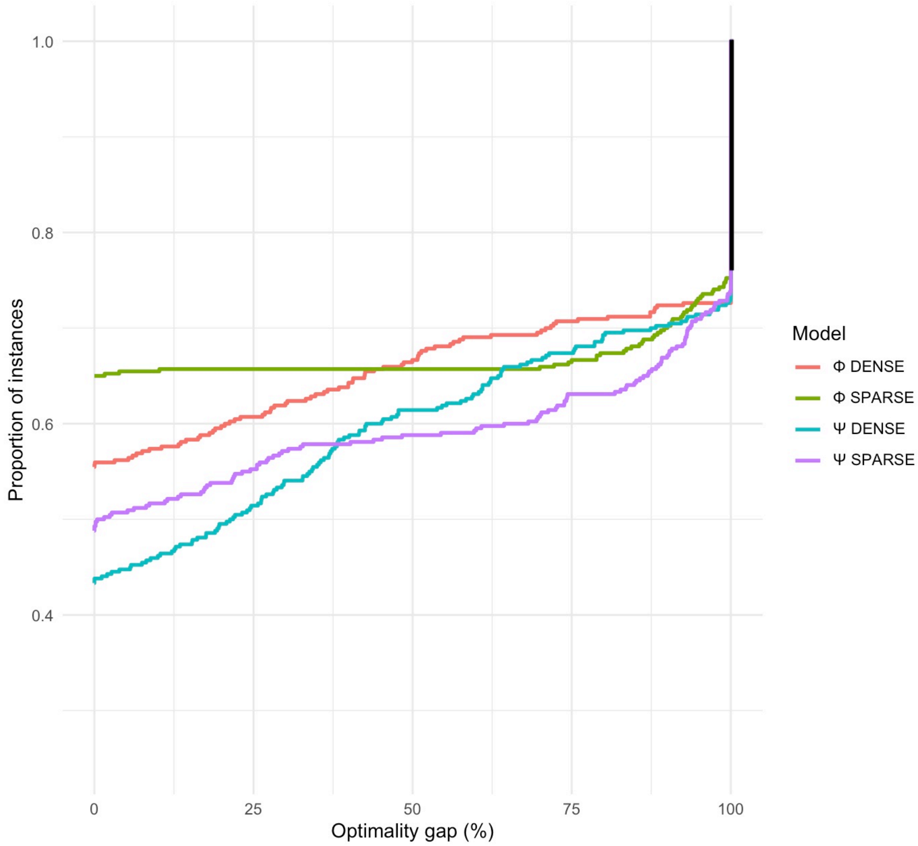


Figure 5: ECDF of optimality gaps ($\texttt{GAP}_{\texttt{opt}}$, proportion) across all instances and objectives under the LOCAL strategy. Higher and leftward curves indicate smaller gaps.

Comparison with van Rossum et al. (2025) approach

To contextualize the performance of our proposed formulations, we compare our results with the recent work of van Rossum et al. (2025), which, to the best of our knowledge, is the only study addressing the TSP-optimal fairness problem under the non-monotonic Range measure. In particular, we benchmark against their *Range branching* algorithm for the RANGE objective, using the TSP column reported in Table 2 of their paper.

Within their branch-and-price framework, van Rossum et al. (2025) adapt the dominance rules in the classical VRP pricing procedure: a partial route dominates another only if it visits the same set of customers with a shorter or equal distance. As noted by the authors, this substantially weakens the dominance mechanism, leading to an almost complete enumeration of feasible routes and making the approach computationally prohibitive beyond 20 customers. Their Table 2 confirms this behavior: for $|V| = 20$, only a few instances are solved to optimality, and for $|V| = 25$, even the root relaxation cannot be solved within the time limit.

We replicate their experimental setting for budgets $B \in \{101\%, 105\%, 110\%\}$ and instance sizes $|V| \in \{15, 20, 25\}$. Table 6 reports the comparison between their TSP-optimal method and our eight model configurations. The computational evidence reveals several notable differences. While the *Range branching* algorithm performs well for $|V| \leq 20$, its scalability is limited. In contrast, our models generate feasible solutions for all instances up to $|V| = 25$ within the prescribed time limit, illustrating the robustness of our approach. Moreover, whereas their method is specifically designed for the range objective, our formulations support a broader class of modeling strategies and equity measures within a unified framework.

It is also important to interpret the optimality gaps in Table 6. For larger instances and tighter budgets, many of our gaps reach values close to 100%. This does not indicate poor solution quality, but rather reflects the fact that the computed lower bounds are often zero when the time limit is reached. In other words, our formulations reliably produce feasible and fair solutions, even when the exact optimality cannot be certified within the allowed computational time.

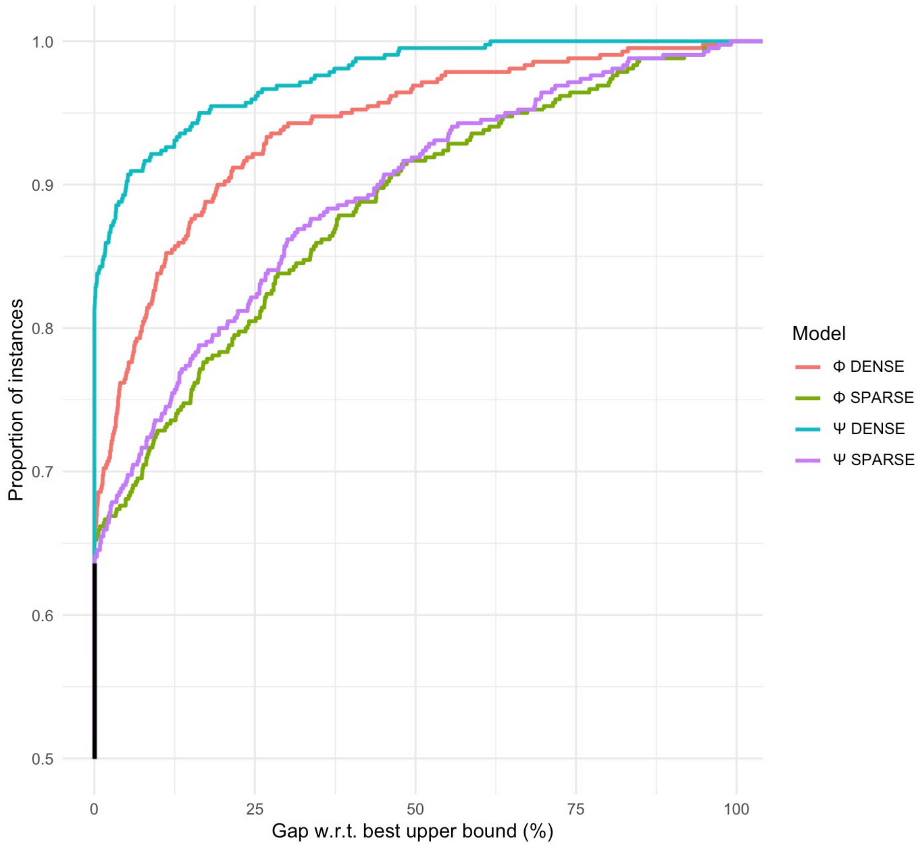


Figure 6: ECDF of gaps to best upper bound ($\texttt{GAP}_{\texttt{upper-best}}$, proportion) across all instances and objectives under the LOCAL strategy. Higher and leftward curves indicate solutions closer to the best known.

Overall, these results show that our exact formulations complement the approach of van Rossum et al. (2025) by extending the range of instance sizes for which feasible and high-quality TSP-optimal fairness solutions can be obtained, providing a rigorous benchmark for future algorithmic developments.

SIZE	$B(\%)$	van Rossum et al. (2025)		Φ MODELS								Ψ MODELS							
				DENSE				SPARSE				DENSE				SPARSE			
				GLOBAL		LOCAL		GLOBAL		LOCAL		GLOBAL		LOCAL		GLOBAL		LOCAL	
		TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP	TIME	GAP
15	101	824.4	0.9	945.0	2.3	926.4	2.7	90.4	**0.0**	71.9	**0.0**	1942.2	13.5	1811.2	15.0	1341.3	4.5	1841.0	7.0
	105	956.9	3.6	877.8	5.2	1168.6	5.8	112.4	**0.0**	66.2	**0.0**	2872.6	30.1	3035.2	36.2	1436.7	6.4	1241.7	4.2
	110	476.6	**0.0**	497.1	**0.0**	571.6	**0.0**	18.0	**0.0**	17.4	**0.0**	2027.4	7.5	2124.5	7.0	122.0	**0.0**	115.2	**0.0**
20	101	3479.4	**10.0**	3605.0	95.3	3606.4	86.5	2366.4	42.8	2619.0	45.9	3610.0	87.9	3608.7	92.5	2757.1	48.1	2708.5	48.5
	105	3600.0	**27.5**	3522.9	85.8	3539.9	89.5	2883.4	52.2	3020.1	49.4	3606.6	82.7	3608.1	87.9	3224.8	55.9	3313.8	56.3
	110	3600.0	41.5	1877.9	27.8	1842.0	21.9	572.8	**0.0**	690.5	**0.0**	2459.5	31.9	2484.0	34.1	2207.5	13.9	2419.1	18.9
25	101	-	-	3673.5	90.2	3586.9	89.7	3623.3	97.6	3607.8	100	3698.5	**86.8**	3611.0	**86.5**	3611.6	100	3636.6	100
	105	-	-	3550.5	89.1	3602.9	93.8	3621.7	100	3617.2	100	3569.3	**87.8**	3546.3	88.4	3616.3	100	3620.8	100
	110	-	-	3608.0	99.0	3604.8	99.4	3632.8	100	3618.0	100	3604.0	**95.3**	3604.5	96.7	3617.3	100	3617.2	100

Table 6: Computational comparison performance for RANGE objective approach of van Rossum et al. (2025).

8 Conclusions

This work develops a unified and flexible framework for modeling fairness in the Vehicle Routing Problem (VRP) by means of ordered objective functions applied to route lengths. By explicitly sorting workloads across vehicles, the proposed approach captures a wide spectrum of equity notions within a single modeling paradigm, ranging from classical monotonic criteria such as mean or min-max to non-monotonic dispersion-based measures including the Gini coefficient, range minimization, variance, and absolute deviations with respect to statistical reference points. This generality enables the design of routing solutions that are not only capacity-feasible and efficient, but also aligned with nuanced fairness considerations that arise in real-world logistics operations.

A key element of the proposed framework is the use of two complementary modeling families to represent ordered fairness objectives. The Φ models rely on explicit ordering constraints that directly enforce the relative position of route lengths in the sorted workload vector. This structure provides a transparent interpretation of fairness and facilitates efficient cut generation at moderate problem sizes. In contrast, the Ψ models are based on k-sum representations, which encode order information implicitly through cumulative sums of the largest workloads.

A central conclusion of this work is that fairness objectives in routing cannot be meaningfully addressed without enforcing internal route efficiency. In particular, non-monotonic equity measures may induce structurally inconsistent solutions if routes are allowed to be suboptimal for their assigned customer sets. To prevent such distortions, we impose per-vehicle TSP-optimality, ensuring that each route is the shortest possible given its allocation of customers. This requirement introduces a tight combinatorial coupling between assignment and routing decisions, which we model through a bilevel optimization framework. The resulting formulation cleanly separates strategic fairness-driven allocation decisions from the tactical realization of efficient routes, making the trade-off between equity and efficiency explicit and operationally interpretable.

From a methodological perspective, the paper demonstrates that value-function reformulations provide a powerful mechanism to embed complex routing subproblems within mixed-integer linear programming models that can be solved exactly. The computational analysis highlights that the effectiveness of the proposed formulations is strongly size-dependent. For small and medium-size instances, sparse Φ formulations combined with local penalty generation offer substantial computational advantages, often reaching optimal solutions orders of magnitude faster than alternative approaches. As instance size increases, however, dense Ψ formulations become more robust, delivering significantly tighter bounds and superior solution quality despite higher computational effort. The experiments further reveal that fairness-oriented non-monotonic objectives are intrinsically more challenging than traditional monotonic metrics. Large optimality gaps persist at higher scales, underscoring the need for tailored formulations and dynamic cut-generation strategies.

Overall, this work represents the first systematic integration of ordered optimization principles into the Vehicle Routing Problem, extending well-established concepts from statistics and facility location to routing contexts. Beyond its immediate contributions, the proposed framework opens several promising research directions, including the development of scalable decomposition algorithms (notably exploiting the column generation procedures), heuristic methods for large-scale instances, and the incorporation of ordered fairness criteria into richer VRP variants with time windows, stochastic demands, or multi-depot structures.

Acknowledgements

The authors acknowledge financial support by grants PID2024-156594NB-C21 (Mathematical optimization for machine learning in complex networks: Foundations and methodological contributions), PID2020-114594GB-C21 (Optimization on data science and network design problems: Large scale network models meet optimization and data science tools), RED2022-134149-T (Thematic Network on Location Science and Related Problems), PCI2024-155024-2 (Optimization over Nonlinear Model Spaces: Where Discrete Meets Continuous Optimization) and IMUS-Maria de Maeztu grant CEX2024-001517-M all funded by Agencia Estatal de Investigación del Ministerio de Ciencia, Innovación y Universidades (MICIU/AEI/10.13039/501100011033).

References

Adams, J. S. (1963). Towards an understanding of inequity. The Journal of Abnormal and Social Psychology, 67(5):422.

Applegate, D., Cook, W., Dash, S., and Rohe, A. (2002). Solution of a min-max vehicle routing problem. INFORMS Journal on Computing, 14(2):132–143.

Archetti, C., Coelho, L., Speranza, M., and Vansteenwegen, P. (2025). Beyond fifty years of vehicle routing: Insights into the history and the future. European Journal of Operational Research.

Artigues, C., Jozefowiez, N., and Sarpong, B. (2018). Column generation algorithms for bi-objective combinatorial optimization problems with a min–max objective. EURO Journal on Computational Optimization, 6(2):117–142.

Balinski, M. L. and Quandt, R. E. (1964). On an integer program for a delivery problem. Operations Research, 12(2):300–304.

Bektaş, T. and Letchford, A. N. (2020). Using lp-norms for fairness in combinatorial optimisation. Computers & Operations Research, 120:104975.

Bertazzi, L., Golden, B., and Wang, X. (2015). Min–max vs. min–sum vehicle routing: A worst-case analysis. European Journal of Operational Research, 240(2):372–381.

Bertsimas, D., Farias, V. F., and Trichakis, N. (2011). The price of fairness. Operations Research, 59(1):17–31.

Bertsimas, D., Farias, V. F., and Trichakis, N. (2012). On the efficiency-fairness trade-off. Management Science, 58(12):2234–2250.

Blanco, V. and Gázquez, R. (2023). Fairness in maximal covering location problems. Computers & Operations Research, 157:106287.

Blanco, V., Marín, A., and Puerto, J. (2024). Intra-facility equity in discrete and continuous p-facility location problems. Computers & Operations Research, 162:106487.

Blanco, V., Pozo, M. A., Puerto, J., and Torrejon, A. (2025). Linear, nested, and quadratic ordered measures: Computation and incorporation into optimization problems. arXiv preprint arXiv:2503.18097.

Boland, N., Domínguez-Marín, P., Nickel, S., and Puerto, J. (2006). Exact procedures for solving the discrete ordered median problem. Computers & Operations Research, 33(11):3270–3300.

Campbell, A. M., Vandenbussche, D., and Hermann, W. (2008). Routing for relief efforts. Transportation Science, 42(2):127–145.

Cherkesly, M., Contardo, C., and Gruson, M. (2025). Ranking decomposition for the discrete ordered median problem. INFORMS Journal on Computing, 37(2):230–248.

Daniell, P. (1920). Observations weighted according to order. American Journal of Mathematics, 42(4):222–236.

Dantzig, G. B. and Ramser, J. H. (1959). The truck dispatching problem. Management Science, 6(1):80–91.

Deleplanque, S., Labbé, M., Ponce, D., and Puerto, J. (2020). A branch-price-and-cut procedure for the discrete ordered median problem. INFORMS Journal on Computing, 32(3):582–599.

Dempe, S. (2002). Foundations of bilevel programming. Springer.

Dempe, S. and Zemkoho, A. B. (2013). The bilevel programming problem: reformulations, constraint qualifications and optimality conditions. Mathematical Programming, 138(1):447–473.

Errami, N., Queiroga, E., Sadykov, R., and Uchoa, E. (2024). VRPSolverEasy: a Python library for the exact solution of a rich vehicle routing problem. INFORMS Journal on Computing, 36(4):956–965.

Fachini, R. F. and Armentano, V. A. (2020). Logic-based benders decomposition for the heterogeneous fixed fleet vehicle routing problem with time windows. Computers & Industrial Engineering, 148:106641.

Feillet, D., Dejax, P., and Gendreau, M. (2005). Traveling salesman problems with profits. Transportation Science, 39(2):188–205.

Gini, C. (1921). Measurement of inequality of incomes. The Economic Journal, 31(121):124–125.

Halvorsen-Weare, E. E. and Savelsbergh, M. W. (2016). The bi-objective mixed capacitated general routing problem with different route balance criteria. European Journal of Operational Research, 251(2):451–465.

Harmati, I. Á., Coroianu, L., and Fullér, R. (2024). The median under orness. Fuzzy Sets and Systems, 481:108901.

Hooker, J. N. and Williams, H. P. (2012). Combining equity and utilitarianism in a mathematical programming model. Management Science, 58(9):1682–1693.

Huang, M., Smilowitz, K., and Balcik, B. (2011). Models for relief routing: Equity, efficiency and efficacy. Procedia - Social and Behavioral Sciences, 17:416–437.

Huber, P. J. (2009). Robust Statistics. John Wiley & Sons, Inc.

Jozefowiez, N., Semet, F., and Talbi, E.-G. (2002). Parallel and hybrid models for multi-objective optimization: Application to the vehicle routing problem. In International conference on parallel problem solving from nature, pages 271–280. Springer.

Kalcsics, J., Nickel, S., Puerto, J., and Tamir, A. (2002). Algorithmic results for ordered median problems. Operations Research Letters, 30(3):149–158.

Karsu, Ö. and Morton, A. (2015). Inequity averse optimization in operational research. European Journal of Operational Research, 245(2):343–359.

Kleinert, T., Labbé, M., Ljubić, I., and Schmidt, M. (2021). A survey on mixed-integer programming techniques in bilevel optimization. EURO Journal on Computational Optimization, 9:100007.

Labbé, M., Ponce, D., and Puerto, J. (2017). A comparative study of formulations and solution methods for the discrete ordered p-median problem. Computers & Operations Research, 78:230–242.

Laporte, G. (2009). Fifty years of vehicle routing. Transportation Science, 43(4):408–416.

Laporte, G. and Louveaux, F. V. (1998). Solving stochastic routing problems with the integer L-shaped method. Springer.

Ljubić, I., Pozo, M. A., Puerto, J., and Torrejón, A. (2024). Benders decomposition for the discrete ordered median problem. European Journal of Operational Research.

López Sánchez, A., Lujak, M., Semet, F., and Billhardt, H. (2024). How to achieve fair and efficient cooperative vehicle routing? AI Communications, 37(2):223–245.

Lozano, J., González-Gurrola, L.-C., Rodríguez-Tello, E., and Lacomme, P. (2016). A statistical comparison of objective functions for the vehicle routing problem with route balancing. In 2016 Fifteenth Mexican international conference on artificial intelligence (MICAI), pages 130–135. IEEE.

Mandell, M. B. (1991). Modelling effectiveness-equity trade-offs in public service delivery systems. Management Science, 37(4):467–482.

Mansini, R., Ogryczak, W., and Speranza, M. G. (2007). Conditional value at risk and related linear programming models for portfolio optimization. Annals of Operations Research, 152(1):227–256.

Marín, A., Ponce, D., and Puerto, J. (2020). A fresh view on the discrete ordered median problem based on partial monotonicity. European Journal of Operational Research, 286(3):839–848.

Marsh, M. T. and Schilling, D. A. (1994). Equity measurement in facility location analysis: A review and framework. European Journal of Operational Research, 74(1):1–17.

Matl, P., Hartl, R. F., and Vidal, T. (2018). Workload equity in vehicle routing problems: A survey and analysis. Transportation Science, 52(2):239–260.

Mesa, J. A., Puerto, J., and Tamir, A. (2003). Improved algorithms for several network location problems with equality measures. Discrete Applied Mathematics, 130(3):437–448.

Nickel, S. and Puerto, J. (2006). Location Theory: A Unified Approach. Springer Science & Business Media.

Ogryczak, W. and Tamir, A. (2003). Minimizing the sum of the k largest functions in linear time. Information Processing Letters, 85(3):117–122.

Pessoa, A., Sadykov, R., Uchoa, E., and Vanderbeck, F. (2020). A generic exact solver for vehicle routing and related problems. Mathematical Programming, 183:483–523.

Pozo, M. A., Puerto, J., and Torrejón, A. (2024). The ordered median tree location problem. Computers & Operations Research, 169:106746.

Puerto, J., Rodríguez-Chía, A. M., and Tamir, A. (2017). Revisiting k-sum optimization. Mathematical Programming, 165:579–604.

Puerto, J. and Torrejon, A. (2025). A fresh view on least quantile of squares regression based on new optimization approaches. Expert Systems with Applications, page 127705.

Rawls, J. (1971). A Theory of Justice. Cambridge, Harvard University Press.

Reiter, P. and Gutjahr, W. J. (2012). Exact hybrid algorithms for solving a bi-objective vehicle routing problem. Central European Journal of Operations Research, 20:19–43.

Savas, E. S. (1978). On equity in providing public services. Management Science, 24(8):800–808.

Svestka, J. A. and Huckfeldt, V. E. (1973). Computational experience with an m-salesman traveling salesman algorithm. Management Science, 19(7):790–799.

Toth, P. and Vigo, D. (2014). Vehicle routing: problems, methods, and applications. SIAM.

Tsang, M. Y. and Shehadeh, K. S. (2025). A unified framework for analyzing and optimizing a class of convex fairness measures. Operations Research.

Uchoa, E., Pecin, D., Pessoa, A., Poggi, M., Vidal, T., and Subramanian, A. (2017). New benchmark instances for the capacitated vehicle routing problem. European Journal of Operational Research, 257(3):845–858.

van Rossum, B., Chen, R., and Lodi, A. (2025). Efficient branching rules for optimizing range and order-based objective functions. Mathematical Programming, pages 1–34.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on Systems, Man, and Cybernetics, 18(1):183–190.