57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Two-echelon routing in last-mile delivery: Formulations and exact methods

Fernando Senna
Departamento de Engenharia de Produção da Universidade Federal de São Carlos (DEP/UFSCar)
Rodovia Washington Luís – km 235, São Carlos-SP
fernandordsenna@gmail.com

Advisor: Reinaldo Morabito, Co-Advisor: Pedro Munari
Departamento de Engenharia de Produção da Universidade Federal de São Carlos (DEP/UFSCar)
Rodovia Washington Luís – km 235, São Carlos-SP
morabito@dep.ufscar.br, munari@dep.ufscar.br

ABSTRACT

Last-mile delivery systems are getting more and more complex with the increase of urban deliveries. To overcome this challenge, several variants of the vehicle routing problem have been proposed. A popular approach is based on two-echelon systems, in which goods are first taken from depots to intermediary points and then shipped to the final customers by smaller vehicles or walking deliverymen. Two classical problems with these characteristics are the two-echelon location-routing problem (2E-LRP) and the vehicle routing problem with time windows and multiple deliverymen (VRPTWMD). This paper provides a summary of a master's dissertation that addresses both problems. For the 2E-LRP, novel mixed-integer programming formulations were proposed, and they were theoretically and computationally compared. Improved best-known lower bounds and new optimal solutions of benchmark instances were found. For the VRPTWMD, variants were introduced to better reflect realistic settings. These variants were formulated and exact methods to solve them were proposed.

KEYWORDS. Vehicle routing. Two-echelon location-routing problem. Multiple deliverymen.

Paper topics. Logistics and Trasportation. Discrete Optimization.

https://proceedings.science/p/212454?lang=pt-br

1. Introduction

Vehicle routing problems (VRPs) are among the most popular combinatorial optimization problems. The most common VRP variant, the capacitated VRP, consists of a set of customers with demands that should be fulfilled by a fleet of vehicles with capacity constraints. The goal of the problem is to define vehicle routes that minimize the delivery costs, while serving all demands and respecting the capacity of the vehicles [Lysgaard et al., 2004]. Many variants of this problem have been introduced in the literature in order to reflect operational constraints, such as heterogeneous fleet, time windows, and pickup and delivery schemes [Golden et al., 2008].

In last-mile delivery systems (with delivery to the final customers), several challenges arise due to poor traffic conditions, city regulations, and the difficulty of finding parking locations. To better accomodate these issues, inventive logistics schemes have been proposed in the literature, such as two-echelon routing. In these settings, larger vehicles ship goods from the depots to intermediary points or parking locations, from which smaller vehicles or walking deliverymen take them to the final customers [Cuda et al., 2015]. Two important combinatorial optimization problems that deal with two-echelon routing are the two-echelon location-routing problem (2E-LRP) and the VRP with time windows and multiple deliverymen (VRPTWMD). This paper provides a summary of a master's dissertation that addresses both of these problems.

The 2E-LRP was introduced by Boccia et al. [2011] and concerns both facility location and vehicle routing in two echelons. Most of the literature uses compact (with a polynomial number of variables and constraints) mixed-integer programming (MIP) formulations based on arc-variables with a vehicle index to represent the problem. This kind of formulation for VRP variants is known to yield poor performance when solved by general-purpose solvers [Furtado et al., 2017]. To overcome this issue, we have proposed three novel compact formulations based on two-index arc variables for the 2E-LRP and introduced valid inequalities to all of them. These formulations have been theoretically compared to evaluate the strength of their linear programming relaxations (LRs). Computational experiments suggest that the novel formulations lead to better solver performance than the benchmark one. They have lead to the improvement of 125 best-known lower bounds and the discovery of 55 new optimal solutions for the 131 benchmark instances evaluated.

The VRPTWMD was originally proposed by Pureza et al. [2012]. In this problem, vehicles work as moving depots that park at specific parking locations, from which the (multiple) deliverymen travelling with the vehicles serve the customers. Including more than one deliveryman in each vehicle speeds up the delivery process, making vehicles more efficient, especially in densely populated urban centers. This leads to an opportunity for cost reduction.

As originally proposed, the VRPTWMD considers two simplifying hypotheses: (*i*) that the customers have been divided into clusters that are served by a single stop of a vehicle in a preprocessing phase, and (*ii*) that the deliveryman routes can be predefined. Hence, the decisions of the problem are the vehicle routes and the number of deliverymen traveling in each vehicle. This has been the most adopted approach in the literature [Álvarez and Munari, 2017].

Senarclens de Grancy and Reimann [2015] were the first to consider the customer clustering as a decision, to which they proposed two heuristics. Later, Senarclens de Grancy [2015] combined the clustering with the vehicle routing. Nevertheless, to the best of our knowledge, the optimization of the deliveryman routes has never been addressed. To bridge this gap, we have introduced two novel problems. The VRPTWMD and two-level routing (VRPTWMD-2R) includes the deliverymen routes in the optimization, while still considering that clustering is previously defined. The VRPTWMD, customer clustering, and two-level routing (VRPTWMD-C2R) includes both the deliveryman routes and the customer clustering in the optimization. The definition of these variants is important because they model a more realistic application, being closer to what is really decided

https://proceedings.science/p/212454?lang=pt-br

in practice. They also lead to further cost reductions. Nevertheless, solving these variants is much more chalenging than solving the VRPTWMD as proposed by Pureza et al. [2012].

Both the VRPTWMD-2R and the VRPTWMD-C2R have been formally defined and formulated. To improve solvability, valid inequalities and Benders reformulations have been proposed for both problems. Computational experiments show how these developments lead to high performing exact algorithms to solve the problems. Furthermore, managerial insights suggest that the inclusion of these decisions in the problem lead to an average 10% cost reduction.

In summary, the contributions of this dissertation are the proposition of novel formulations and exact solution methods for two-echelon VRP variants that arise in last-mile delivery. In a context of growing urban deliveries, it is of major importance that the operations research literature discusses realistic combinatorial optimization problems that can be applied in modern settings. It is also essential to develop high-performance solution methods for these problems. The dissertation summarized in this article contributes to this discussion.

The remainder of this document is organized as follow. In Section 2, the developments regarding the 2E-LRP are presented. Section 3 discusses the VRPTWMD-2R. Section 4 introduces the discussion about the VRPTWMD-C2R. In Section 5, the master's dissertation development and resulting publications are outlined. Finally, Section 6 presents concluding remarks.

2. The two-echelon location-routing problem

The 2E-LRP is a combinatorial optimization problem that combines both location and routing decisions in a two-echelon scheme. Let $\mathcal{E} = \{1, 2\}$ represent the set of echelons, with $e = 1$ meaning the first echelon (FE) and $e = 2$ the second echelon (SE). Let $\mathcal{N}^e$ and $\mathcal{A}^e$ be the sets of nodes and arcs, respectively, of echelon $e \in \mathcal{E}$. Hence, the 2E-LRP can be defined over a graph $G = (\mathcal{N}, \mathcal{A})$, with $\mathcal{N} = \mathcal{N}^1 \cup \mathcal{N}^2$ and $\mathcal{A} = \mathcal{A}^1 \cup \mathcal{A}^2$. Figure 1 presents a solution for an instance of the problem with three potential platforms (of which only one is used), three potential satellites (of which two are used), and four customers. The goods are initially placed in the platform, from which they are shipped to the satellites by an FE vehicle. Then, two SE vehicles take these goods to the customers.

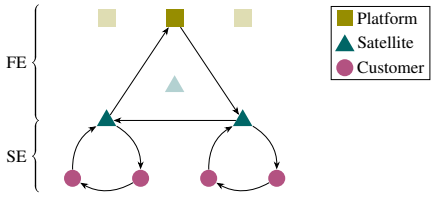


Figure 1: An illustrative example of the 2E-LRP.

The decisions of the problem are the facilities (satellites and platforms) to open and the vehicle routes. The goal is to minimize costs comprising of fixed costs of opening facilities and using vehicles, and distance-related costs of the vehicle routes. Constraints such as respecting vehicle and facility capacities, and serving all customer demands must be satisfied.

When introducing the 2E-LRP, Boccia et al. [2011] proposed two compact formulations, and their experimental results showed that the one that leads to the best solver performance is based on arc variables with a vehicle index: x^e_{ijk}, with $(i, j) \in \mathcal{A}^e, k \in \mathcal{K}^e, e \in \mathcal{E}$, with $\mathcal{K}^e$ representing the set of vehicles available in echelon e. Thus, most of the literature that followed, adopted variations of this formulation to represent the 2E-LRP and its variants [Cuda et al., 2015]. We shall refer to this formulation as CF1.

Despite this original result, it is commonly observed in other VRP variants that this kind of variable leads to poor solver performance [Furtado et al., 2017]. The difficulty with defining

https://proceedings.science/p/212454?lang=pt-br

compact formulations for the 2E-LRP without a vehicle index in the variables is the need to have the vehicle routes ending where they started. We have introduced three formulations that overcome this issue and are based on two-index arc variables: x_{ij}^e, with $(i,j) \in \mathcal{A}^e, e \in \mathcal{E}$.

Formulation CF2 relies on binary assignment variables which associate destinations (satellites in the FE and customers in the SE) with origins (platforms in the FE and satellites in the SE). A lifting of this formulation is referred to as ICF2. The third formulation (CF3) we have introduced is based on continuous variables that indicate the origin node of the vehicle visiting each satellite (FE) or customer (SE). Moreover, we have adapted CF1 to rely on commodity flow constraints [Gavish and Graves, 1978] to make comparison fairer with CF2, ICF2, and CF3, since these three formulations rely on this type of constraints. This formulation is referred to as ICF1. We have also introduced several valid inequalities for the formulations.

We have theoretically compared these formulations based on the strength of their linear programming relaxations (LRs), as disclosed in Proposition 1. Let P and Q be two formulations and let $F(\mathrm{P}_L)$ represent the feasible region of the LR of P (P_L). If P and Q are defined over the same variable space and $F(\mathrm{P}_L) \subsetneqq F(\mathrm{Q}_L)$, P has a stronger LR than Q. If the variable spaces are different, P has a stronger LR than Q if there is a linear transformation T such that $T(F(\mathrm{P}_L)) \subseteq F(\mathrm{Q}_L)$ and there does not exist a linear transformation T' such that $T'(F(\mathrm{Q}_L)) \subseteq F(\mathrm{P}_L)$. Formulations with stronger LRs tend to yield better solver performance. If neither P has a stronger LR than Q nor Q has a stronger LR than P, they are not comparable.

Proposition 1. *The following relationships between LRs are true:*

(a) The LRs of CF1 and ICF1 are not comparable;
(b) CF2 has a stronger LR than ICF1;
(c) The LRs of CF2 and CF3 are not comparable;
(d) CF3 has a stronger LR than ICF1;
(e) ICF2 has a stronger LR than ICF1;
(f) ICF2 has a stronger LR than CF2;
(g) ICF2 has a stronger LR than CF3.

Computational experiments confirm the impact that these theoretical results have in the performance of a general-purpose solver. The three novel formulations defined with two-index arc variables lead to better solver performance than those based on variables with a vehicle index. CF3 is the best performing formulation because it combines a tight bound with a small number of variables and constraints. The valid inequalities also provide improvements to the solver performance. Furthermore, our experiments have lead to the improvement of 125 best-known lower bounds and 55 new optimal solutions for the 131 benchmark instances evaluated, pushing forward the state of the art of the algorithms to solve the 2E-LRP.

3. The VRPTWMD and two-level routing

As discussed in Section 1, the VRPTWMD has never been studied with the deliveryman routes included in the optimization. This setting works well if the deliveryman capacities are small compared to the customer demands and, hence, the only feasible routes are back and forth trips from the vehicles to the customers. When this is not the case, preprocessing the deliveryman routes as proposed by Pureza et al. [2012] does not work well. To overcome this issue, we have introduced the VRPTWMD-2R, which optimizes the deliveryman routes conjointly with the vehicle routes and the number of deliveryman traveling with each vehicle. This problem is a more realistic variant of the VRPTWMD. It is also much more difficult to solve since the number of feasible solutions greatly increases.

Figure 2 illustrates the problem. In Figure 2a, an instance of the VRPTWMD-2R is presented, with four clusters of customers and their respective parking locations. Figure 2b shows a

https://proceedings.science/p/212454?lang=pt-br

feasible solution for this instance, with a vehicle with two deliverymen serving the clusters on the right hand-side of the picture and a vehicle with a single deliverymen serving the other ones.



(a) Instance. (b) Solution.

Figure 2: An illustrative example of the VRPTWMD-2R.

We have formally defined the VRPTWMD-2R and proposed a MIP formulation for it. The MIP formulation has a weak LR that leads to poor performance of general-purpose solvers. Therefore, we have introduced several valid inequalities and a Benders reformulation for the problem, which is solved by a branch-and-Benders-cut (BBC) algorithm. To further enhance the algorithms, several MIP problems are solved in a preprocessing phase, leading to the definition of tighter valid inequalities and a better defined master problem for the BBC.

Our computational experiments show the positive impact of the proposed valid inequalities, with the number of optimal solutions found by the solver going from 11 to 90 with their inclusion. Moreover, the adoption of the BBC leads to an even better performance, finding 129 optimal solutions for the 145 instances evaluated, many of them of realistic size. Finally, managerial insights suggest that the inclusion of the decision on the deliveryman routes in the optimization problem leads to an average of 10% cost reduction in the optimal solution. These savings can be further enhanced with the adoption of a higher limit in the number of deliverymen in each vehicle or faster deliverymen, such as bicycles.

4. The VRPTWMD, customer clustering and two-level routing

To further our analysis of the VRPTWMD, we have introduced the VRPTWMD-C2R, which combines both deliveryman routes and customer clustering in the optimization problem. Compared to the VRPTWMD-2R, the amount of feasible solutions grow exponentially, with the difficulty in solving the problem increasing significantly. Nevertheless, this problem is much more realistic, leading to a proper evaluation of the decisions to be made in practice.

We have formally defined the VRPTWMD-C2R and proposed a MIP formulation to it. To strengthen the formulation, we have introduced more than 20 valid inequalities. A Benders reformulation was proposed and solved by means of a BBC. Nevertheless, due to the increased difficulty of the problem, further developments were necessary for the good performance of the BBC. First, several lower bounds and properties were introduced and used to strengthen the Benders reformulation. Let us denote by N^1 the set of parking locations, by N^2 the set of customers, by Q^1 and Q^2 respectively the load capacities of vehicles and deliverymen, by q^2_h the demand of customer h, by s^2_h the service time needed in customer h, by t^2_{ih} the travel time of a deliveryman to go from node i to node h, by t^1_{ij} the travel time of a vehicle to go from a node i to a node j, by z_{ih} a binary variable that indicates whether customer h is assigned to parking location i, by M_L the maximum number of deliverymen that can travel in a vehicle, by c^2 the unitary distance cost of deliveryman routes, and by d^2_{ih} the distance between nodes i and h. The properties and lower bounds are defined in Propositions 2 to 6.

Proposition 2. *If the triangular inequality holds, there is an optimal solution in which only parking locations with customers assigned to them are visited.*

Proposition 3. *If the triangular inequality holds, a lower bound on the time spent in a parking location $i \in N^1$ visited by l deliverymen is given by* $\sum_{h \in N^2} \left(s_h^2 + (t_{ih}^2 + t_{hi}^2) \frac{q_h^2}{Q^2} \right) \frac{z_{ih}}{l}$.

Proposition 4. *If the triangular inequality holds, a lower bound on the total travel time of the vehicles is* $\frac{1}{Q^1} \sum_{i \in N^1} \sum_{h \in N^2} (t_{0i}^1 + t_{i0}^1) q_h^2 z_{ih}$.

Proposition 5. *If the triangular inequality holds, a lower bound on the time the vehicles stay out of the depot is* $\frac{1}{Q^1} \sum_{i \in N^1} \sum_{h \in N^2} (t_{0i}^1 + t_{i0}^1) q_h^2 z_{ih} + \frac{1}{M_L} \sum_{h \in N^2} \left(s_h^2 + \frac{1}{Q^2} \sum_{i \in N^1} (t_{ih}^2 + t_{hi}^2) q_h^2 z_{ih} \right)$.

Proposition 6. *If the triangular inequality holds, a lower bound on the cost of the deliveryman routes inside a cluster $i \in N^1$ is* $c^2 \sum_{h \in N^2} (d_{ih}^2 + d_{hi}^2) q_h^2 z_{ih}$.

On top of these theoretical developments, computational improvements were also necessary to enhance the algorithms. Due to the large amount of potential clusters, there are many subproblems solved by the BBC that do not provide useful cuts. Hence, we proposed a two-phase BBC (2P-BBC), in which, in the first phase, few cuts are introduced, until a gap plateau (when there is little improvement in both the lower and upper bounds) is reached. After that, all cuts are included. This avoids the inclusion of an excessive and unnecessary number of cuts at the beginning of the procedure, when they are not useful for the solution algorithm. This also reduces the time spent solving the subproblem. The drawback of this approach is that, in the first phase, the algorithm is unable to find good feasible solutions. To overcome this issue, we have proposed a MIP heuristic that provides a warm start for the BBC, improving the algorithm performance.

Computational experiments show the suitability of the proposed methodology to solve the problem. The VRPTWMD-C2R formulation has a very weak LR, which is significantly improved by the valid inequalities (the provided bound is over 25 times higher on average). Moreover, the 2P-BBC coupled with the MIP heuristic leads to major improvements in the algorithmic performances.

Finally, managerial insights suggest that, although the inclusion of the deliveryman routes in the optimization is always beneficial, the customer clustering depends on the instances characteristics. In instances with parking locations far from each other and customers closely grouped around them, clustering can be effectively preprocessed. Nevertheless, when there are many potential parking locations close one another with customers around them, the inclusion of the clustering decision in the optimization leads to great opportunities for cost reduction.

5. Dissertation development and publications

During his master's, the candidate has achieved the highest grades in all courses he took. Moreover, he has had the opportunity to develop part of his research abroad. He has been a research intern at the Centre interuniversitaire de recherche sur les reseaux d'entreprise, la logistique et le transport (CIRRELT) at Université Laval, in Quebec, Canada, under the supervision of Prof. Dr. Leandro C. Coelho. This collaboration was essential for the candidate development, with the research conducted during this period being a central part of this dissertation. He has also been a visiting researcher at the University of Edinburgh Business School, under the supervision of Prof. Dr. Douglas Alem. The collaboration with Prof. Alem was also very fruitful, but this work ended up not being included in the dissertation because it was not related to VRPs.

The candidate has also participated in several scientific events during his master's. In 2022, he presented his work at the Escuela Latinoamericana de Verano en Investigación Operativa (ELAVIO) in Monterrey, Mexico; and at the Simpósio Brasileiro de Pesquisa Operacional (SBPO) in Juiz de Fora, Brazil. He also took part in the 8$^{\text{th}}$ Workshop on Mathematical Solutions in São

https://proceedings.science/p/212454?lang=pt-br

Carlos, Brazil, organized by the Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI) in 2022. In 2023, he presented his work at the joint conference of the Canadian Operations Research Society (CORS) with Journées de l'Optimisation (JOPT) in Montreal, Canada; at the conference of the International Federation of Operational Research Societies (IFORS) in Santiago, Chile; and at the SBPO in São José dos Campos, Brazil. In 2024, he gave a presentation at the Oficina Nacional de Problemas de Corte, Empacotamento, Planejamento e Programação da Produção e Correlatos (ONPCE), in Bauru, Brazil. All these events were essential for the candidate development and networking.

Finally, the developed research has lead to three journal papers, with one of them already published at the European Journal of Operational Research and the other two under revision in renowned journals. A conference paper and an extended abstract are also part of the resulting publications of this master's dissertation. The full list is presented below:

Paper published in scientific journal

- *An exact method for a last-mile delivery routing problem with multiple deliverymen*. Published in the European Journal of Operational Research. This paper introduces the VRPTWMD-2R and the methods discussed in Section 3.

Papers under review in scientific journals

- *The two-echelon location-routing problem: A comparative analysis of novel and existing compact formulations*. Under review at Transportation Research Part E: Logistics and Transportation Review. This paper discusses the developments for the 2E-LRP presented in Section 2.
- *Last-mile delivery with multiple deliverymen: Formulation and exact solution methods for a two-echelon vehicle routing problem*. Under review at the European Journal of Operational Research. This paper reflects the VRPTWMD-C2R and the discussion of Section 4.

Conference paper

- *O problema de roteamento de veículos com janelas de tempo, múltiplos entregadores e dois níveis de decisão*. Published in the proceedings of the 54th Simpósio Brasileiro de Pesquisa Operacional.

Extended abstract

- *Roteamento de veículos com múltiplos entregadores envolvendo clusterização e dois níveis de rotas com aplicações em last-mile delivery*. Published in the proceedings of the 55th Simpósio Brasileiro de Pesquisa Operacional.

6. Conclusion

In this paper, we have presented a summary of a master's dissertation that addresses two important combinatorial optimization problems in last-mile delivery: the two-echelon location-routing problem and the vehicle routing problem with time windows and multiple deliverymen. For the 2E-LRP, we proposed three novel formulations based on two-index arc variables and several valid inequalities. These formulations were theoretically compared by means of the strength of their LRs. Computational experiments showed that the novel formulations lead to much better solver performance than the benchmark one. Moreover, 125 best-known lower bounds were improved and 55 new optimal solutions were found for the 131 benchmark instances evaluated.

Regarding the VRPTWMD, two novel variants were proposed: the VRPTWMD-2R and the VRPTWMD-C2R. The first includes the decision on which deliveryman routes to perform in the optimization, while the latter include both the deliveryman routes and the customer clustering in the optimization problem. Both of these decisions were usually overlooked by the literature, which assumed that they could be precomputed. However, our developments indicate that the new variants

not only model more realistic settings but they also lead to further cost reductions compared to the original VRPTWMD.

We have formally defined both variants and proposed MIP formulations for them. These formulations were strengthened by several valid inequalities. We introduced Benders reformulations, which were solved by BBC algorithms. For the VRPTWMD-C2R, analytical lower bounds were proposed and improvements for the BBC were discussed. Computational experiments show the suitability of the proposed methodology to solve these problems. Finally, managerial insights suggest that the inclusion of both decisions in the optimization lead to significant cost reductions.

In conclusion, the referred dissertation provides significant contributions for the state-of-the-art surrounding VRP problems applied to last-mile delivery. With the expansion of urban centers, last-mile delivery has become a major challenge for logistics service providers. The presented developments contribute to the knowledge of combinatorial optimization problems, MIP formulations, and exact solution methods in this area. The resulting publications highlight the impact of the research results.

References

Álvarez, A. and Munari, P. (2017). An exact hybrid method for the vehicle routing problem with time windows and multiple deliverymen. *Computers & Operations Research*, 83:1–12.

Boccia, M., Crainic, T. G., Sforza, A., and Sterle, C. (2011). Location-routing models for designing a two-echelon freight distribution system. Technical report CIRRELT-2011-06, Université de Montréal.

Cuda, R., Guastaroba, G., and Speranza, M. G. (2015). A survey on two-echelon routing problems. *Computers & Operations Research*, 55:185–199.

Furtado, M. G. S., Munari, P., and Morabito, R. (2017). Pickup and delivery problem with time windows: A new compact two-index formulation. *Operations Research Letters*, 45(4):334–341.

Gavish, B. and Graves, S. C. (1978). The traveling salesman problem and related problems. Technical report, Operations Research Center, Massachusetts Institute of Technology. OR 078-78.

Golden, B., Raghavan, S., and Wasil, E. (2008). *The vehicle routing problem: Latest advances and new challenges*. Springer US, Nova York, EUA.

Lysgaard, J., Letchford, A. N., and Eglese, R. W. (2004). A new branch-and-cut algorithm for the capacitated vehicle routing problem. *Mathematical Programming*, 100:423–445.

Pureza, V., Morabito, R., and Reimann, M. (2012). Vehicle routing with multiple deliverymen: Modeling and heuristic approaches for the VRPTW. *European Journal of Operational Research*, 218(3):636–647.

Senarclens de Grancy, G. (2015). An adaptive metaheuristic for vehicle routing problems with time windows and multiple service workers. *Journal of Universal Computer Science*, 21(9):1143–1167.

Senarclens de Grancy, G. and Reimann, M. (2015). Evaluating two new heuristics for constructing customer clusters in a VRPTW with multiple service workers. *Central European Journal of Operations Research*, 23(2):479–500.

https://proceedings.science/p/212454?lang=pt-br