A concise guide to existing and emerging vehicle routing problem variants

Thibaut Vidal*
Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Brazil
vidalt@inf.puc-rio.br

Gilbert Laporte
CIRRELT, Canada Research Chair in Distribution Management and HEC Montréal, Canada
gilbert.laporte@cirrelt.ca

Piotr Matl
Department of Business Decisions and Analytics, University of Vienna, Austria
piotr.matl@univie.ac.at

Abstract. Vehicle routing problems have been the focus of extensive research over the past sixty years, driven by their economic importance and their theoretical interest. The diversity of applications has motivated the study of a myriad of problem variants with different attributes. In this article, we provide a concise overview of existing and emerging problem variants. Models are typically refined along three lines: considering more relevant objectives and performance metrics, integrating vehicle routing evaluations with other tactical decisions, and capturing fine-grained yet essential aspects of modern supply chains. We organize the main problem attributes within this structured framework. We discuss recent research directions and pinpoint current shortcomings, recent successes, and emerging challenges.

Keywords. Transportation, Combinatorial optimization, Vehicle routing problem, Challenges and perspectives

* Corresponding author
1 Introduction

Vehicle routing problems (VRPs) have been the subject of intensive and fast-growing research over sixty years. This is due to their economic importance and their theoretical interest. Using efficient vehicle routes provides a direct competitive advantage to transportation companies, which usually operate with limited profitability margins. Moreover, the fact that these problems share a simple yet rich structure, generalizing the traveling salesman problem, has helped to elevate the VRP family into one of the main testbeds for studies in combinatorial optimization and heuristics. The VRP family can be seen as combinatorial in two senses: 1) because of the number of possible solutions, which grows exponentially with the size of the instances, and 2) because the number of conceivable problem variants also grows exponentially with the variety of problem attributes, i.e., the specific constraints, decision sets and objectives arising from real applications (Vidal et al. 2013).

The VRP research landscape has dramatically evolved over the past two decades. Up to the early 2000s, most methodological studies were centered around a limited subset of operational problems with attributes such as time windows, multiple depots, multiple periods, and heterogeneous fleets. Since then, the number of problem variants has grown rapidly, reflecting the diversity of applications. Vehicle routing algorithms are no longer used only to produce daily routes but also serve as evaluation tools for other strategic and tactical decisions such as facility location, fleet sizing, production, and inventory management (Andersson et al. 2010, Hoff et al. 2010).

The goal of this article is to draw a succinct picture of current research in the field of vehicle routing. It is addressed to researchers and practitioners who wish to consult a concise review of existing problem features and applications. We discuss within a structured framework the main problem attributes and research directions in the field of vehicle routing as of 2020, pinpointing current shortcomings, recent successes and emerging challenges. Given the breadth of the field, a description of every available study is now impractical. This paper therefore does not claim to be exhaustive in its coverage. Instead, we have opted for a structure based on themes rather than on VRP variants, as is the case of several books or review papers, and refer to the books of Golden et al. (2008) and Toth and Vigo (2014) for a more detailed coverage of specific problem variants. This work is organized according to application-centered goals and concerns. From a high-level perspective, a VRP model can be extended along three main lines: 1) considering relevant side metrics, objectives, or combinations of objectives; 2) integrating routing optimization with other business decisions; 3) progressing toward more precise and fine-grained models.

We discuss the academic problem variants and studies according to these three classifications in Sections 2 to 4. Then, we highlight some important challenges and conclude in Section 5.
2 Emerging Objectives – Measuring as a Step Toward Optimizing

Measurement and quantification are central to any optimization algorithm for business processes. Most of the VRP literature considers cost as the main objective, but this does not capture all relevant performance criteria and metrics arising in practice, and many solutions based on cost optimization alone turn out to be impossible to apply in practice. In these contexts, other metrics must be considered, either as additional objectives or as constraints. We subdivide these metrics into seven main categories:

1) **profitability**: performance ratios, profits, outsourcing;
2) **service quality**: cumulative objectives, inconvenience measures, service levels;
3) **equity**: workload balance, service equity, collaborative planning;
4) **consistency**: temporal, person-oriented, regional, or delivery consistency, inconsistency;
5) **simplicity**: compactness, separation, navigation complexity;
6) **reliability**: expected cost or loss, probability of failure;
7) **externalities**: emissions, safety risks.

This section will discuss each of these criteria and the related VRP variants. For some applications, multiple criteria may appear as objectives (using a weighted sum, hierarchical or multi-objective formulation) or as constraints. We analyze how each criterion has been integrated in academic problems, citing key methodological contributions and case studies.

2.1 Profitability

It is safe to say that profitability or cost optimization is the primary concern in the overwhelming majority of VRP studies. Most articles consider the minimization of total routing costs, which may include a fixed cost per route (e.g., vehicle cost, insurance, daily wages) as well as variable costs proportional to distance or travel duration (e.g., fuel consumption, maintenance costs, hourly wages). Moreover, as outlined below, profitability also extends beyond operational costs.

**Performance ratios.** In some situations, the optimization of routing costs is not meaningful and can even be counter-productive if it is not balanced with other performance measures. Especially for problems posed on a rolling horizon, there is a need to consider short-term surrogate objectives that approximate long-term performance goals. A practical example is the class of inventory-routing problems, for which several authors have emphasized the need to optimize the logistic ratio. This is the ratio of routing cost to delivered quantity over the planning horizon (Song and Savelsbergh 2007, Benoist et al. 2011, Archetti et al. 2017b), which measures the average cost to deliver one unit of product. This objective prevents myopic behavior that could arise from pure cost minimization (Archetti et al. 2017b). Another practical example can be found in mobility-on-demand services and in the maximization of the occupancy rate, i.e., the ratio of total passenger travel times to total vehicle travel times (Garaix et al. 2011). VRPs with other fractional objectives, such as profit over time, have been studied in Baldacci et al. (2018).
Profit. Cost minimization often competes with profit maximization in tactical business decisions. This is especially true when the optimizer has the authority to select some of the deliveries, giving rise to the class of VRPs with profits (Archetti et al. 2014b). In most of these problems, customers are associated with individual prizes, and the objective is to maximize the total profit as the difference between collected prizes and routing costs. Other problem variants maximize profit subject to distance or time constraints. These problems are connected to numerous applications in production planning and logistics (Aksen et al. 2012), manufacturing (Lopez et al. 1998), military reconnaissance (Mufalli et al. 2012), and the design of tourist itineraries (Vansteenwegen and Souffriau 2009), among others.

Outsourcing. To respond to growing delivery volumes while limiting the impact of high variance in shipping patterns, many freight forwarding companies regularly outsource a portion of their business to subcontractors. This practice has led to the VRP with private fleet and common carrier (see, e.g., Côté and Potvin 2009), which can be viewed as a special case of VRP with profits in which each customer’s prize represents its outsourcing cost. Several variants of this problem have recently been studied. Krajewska and Kopfer (2009) discuss the impact of considering heterogeneous subcontractors and distinguish three types of direct outsourcing cost: per tour, flat rate per day, and flow-based depending on distance and weight. Ceschia et al. (2011) consider a cost function with coefficients that depend not only on distance and load but also on geographic aspects relating to the most distant customer on a tour. Stenger et al. (2013b) solve a variant with multiple depots, while Stenger et al. (2013a), Gahm et al. (2017) and Dabia et al. (2019) consider nonlinear cost functions arising from volume discounts. Finally, Goeke et al. (2019a) design a state-of-the-art branch-and-price algorithm.

2.2 Service Quality

Although operational efficiency is important, providing superior service quality helps businesses to differentiate themselves from their competitors. Furthermore, profit measures are not the primary concern in some contexts, such as humanitarian relief operations, public transportation, and home healthcare logistics.

Cumulative objectives. In the cumulative VRP (Ngueveu et al. 2010a, Silva et al. 2012), the classical total cost objective is replaced with the sum of individual arrival times at the customers. Objectives of this type can be seen as more service-focused, and they are often proposed as relevant optimization criteria for relief effort operations (Campbell et al. 2008, Golden et al. 2014). The components of a cumulative objective can also be weighted in order to further bias the route structure toward a customer-centered perspective (Huang et al. 2012). In general, cumulative objectives are more appropriate when the distribution of the arrival times, travel times, transported load, etc. is more important than their sum.
Inconvenience measures. Service quality is particularly important for passengers transportation. Common examples include the planning of school bus routes (Park and Kim 2010) and dial-a-ride services organized by home healthcare providers (Cordeau and Laporte 2007). Since service quality is multi-dimensional, many criteria have been proposed to measure its different aspects (Paquette et al. 2012). From an optimization perspective, it is common to introduce measures of customer inconvenience, for example by minimizing the maximum ride time; the maximum time loss defined as the difference between the ride time and the corresponding shortest possible ride; or the deviation from a preset time window in case of earliness or lateness. Importantly, the target levels of these criteria may vary for different sets of customers. For example, emerging mobility-on-demand platforms aim to satisfy different service quality thresholds for different customer segments (e.g., business, standard, budget) (Beirigo et al. 2019).

Service levels. As a general trend worldwide, logistics activities are being increasingly outsourced to third-party logistics (3PL) service providers. Due to large volumes and unforeseen events, 3PL providers can rarely service all requests. To guarantee a certain service level, most 3PL providers establish contracts that stipulate a minimum ratio of on-time deliveries. This gives rise to the VRP with service-level constraints (Bulhões et al. 2018a). In this problem, the deliveries are partitioned into groups, and a minimum percentage of the deliveries (or delivery load) must be fulfilled for each group. Orlis et al. (2019) describe an application to automated teller machine replenishment. In this study, the service levels are treated as soft constraints, and there are penalties for non-fulfillment. Service fulfillment can even be the primary optimization objective in some home healthcare applications (Rasmussen et al. 2012).

2.3 Equity

Efficient solutions are not necessarily equitable. Their acceptance and implementation may be contingent on a sufficiently fair distribution of resources, responsibilities, and benefits among different stakeholders. These concerns have led to a variety of equity criteria in the routing literature, reviewed in Balcik et al. (2011) and Matl et al. (2018, 2019).

Workload balance. In routing problems in the private sector, the most common equity considerations concern internal stakeholders, i.e., the drivers or other personnel providing the service. The aim is to balance the workload allocation in order to ensure acceptance of operational plans, to maintain employee satisfaction and morale, to reduce overtime, and to reduce bottlenecks in resource utilization. Practical examples include balancing the workload of service technicians (Blakeley et al. 2003), home healthcare professionals (Liu 2013), and volunteers (Goodson 2014). Balancing criteria also appear in periodic settings (Groër and Golden 2009, Mendoza et al. 2009) and in tactical planning problems such as service territory design (Butsch et al. 2014, Kalc scs 2015). The workload $W_r$ of a route or service unit is usually operationalized through its total service duration, total demand, number of customers, or some combination of these metrics. The degree of balance is then quantified by applying an inequality func-
tion to the vector of workloads, the most common functions being minimization of the maximum workload \(\min\{\max\{W_{r}\}\}\) and the minimization of the range \(\min\{\max\{W_{r}\} - \min\{W_{r}\}\}\). Care should be taken when modeling equity criteria, because certain combinations of workload metrics and equity functions are not appropriate for guiding mathematical optimization methods. In particular, equity functions that are not monotonic with respect to all workloads (e.g., the range or standard deviation) can lead optimization methods to unnecessarily increase the workload (e.g., longer distance or time) of every route in an effort to artificially satisfy the ill-posed equity criterion (Matl et al. 2018).

**Service equity.** In contrast to private and profit-oriented logistics businesses, public and nonprofit organizations also generally have an obligation of equitable service provision to their external stakeholders, i.e., the users or customers (Balcik et al. 2011). The most common application areas are public transportation and humanitarian logistics. There exists a close connection between the service quality measures discussed in the previous section and service equity. In fact, the min-max constraints on service quality can also be interpreted as equity constraints. However, as discussed and analyzed by Huang et al. (2012) in the context of disaster relief, there can be discrepancies between efficacy (quality of coverage) and equity, as these issues may concern different dimensions, e.g., the quantity of supplies may satisfy the full demand at all the service points while the timeliness of delivery may be very inequitable, or vice versa. Moreover, we note that unless the quality of the worst service is tightly constrained, satisfying the corresponding min-max constraint does not imply an equitable distribution of service quality. To date, these issues have received limited attention in the VRP literature.

**Collaborative planning.** Due to strong competitive pressures and falling profit margins in the logistics sector, carriers have an incentive to form horizontal collaborations that pool their capacities and increase their overall efficiency (Cruijssen et al. 2007, Gansterer and Hartl 2018). In such coalitions, the planning of logistics operations is performed jointly through the exchange or consolidation of transportation requests and a redistribution of costs or gains, which leads to problems of fair resource allocation and profit sharing (Guajardo and Rönqvist 2016, Padilla Tinoco et al. 2017). The collaboration should be stable in the sense that each partner’s individual cost is reduced by joining the partnership and the benefit of these reductions is fairly distributed. Since many of the proposed cost allocation methods relate the distributed cost or gain to the contributed resources (Guajardo and Rönqvist 2016), the routing decisions help determining the achievable savings and the fairest benefit distribution.

### 2.4 Consistency

Cost-optimal routing plans may turn out to be of limited value if they vary too much over time. Customers appreciate being served by familiar faces at regular intervals; service providers are more effective and can personalize their service when they know their customers’ requirements and preferences; drivers are more efficient and drive more safely when they are familiar with the
peculiarities of their routes (Kovacs et al. 2014a). Establishing and maintaining these aspects of familiarity requires routing and service plans to be consistent with respect to various metrics over multiple time periods.

Temporal consistency. One aspect of consistency concerns the timing of the service provision to individual customers. The aim is to provide service at roughly the same time of day and at regular intervals. Initial studies by Groër and Golden (2009), Tarantilis et al. (2012), and Kovacs et al. (2014a) handle this feature by imposing a maximum difference between the latest and earliest arrival times at any customer location. The resulting consistent VRP (ConVRP) is often solved by metaheuristics, since the time constraints create route interdependencies which pose considerable challenges for exact solution approaches (Goeke et al. 2019b). Feillet et al. (2014) suggest an alternative approach that discretizes the day into disjoint time segments and imposes consistency by bounding the number of different time segments during which a customer is served. Some recent works propose self-imposed time windows, whereby the service provider selects for each customer a fixed time window before the demand is known, communicates this information to the customer, and subsequently generates routing plans respecting these commitments during the planning horizon (Jabali et al. 2015, Spliet and Gabor 2015). Other authors have set minimum and maximum time intervals between consecutive customer visits in periodic settings (Gaudioso and Paletta 1992, Coelho and Laporte 2013).

Person-oriented consistency. Another form of consistency relates to the assignment of drivers to customers. If the same driver regularly visits the same customers, the quality and efficiency of the customer service improves as personal relationships become established and the driver becomes more familiar with the customers (Smilowitz et al. 2013). This type of consistency is particularly important in home healthcare logistics, where the quality of the service depends on the nurses’ knowledge of the preference and needs of their patients (Eveborn et al. 2006). Personal consistency is often handled at the tactical level by creating a fixed route for each driver (Christofides 1971, Beasley 1984) or a set of template routes that are adjusted into daily operational routes (Groër and Golden 2009, Tarantilis et al. 2012, Kovacs et al. 2014b). More flexible alternatives focus on directly maximizing the number of times a unique driver visits each customer (Haughton 2007, Smilowitz et al. 2013), ensuring that each driver visits at least a certain fraction of their assigned customers (Spliet and Dekker 2016) or bounding the number of different drivers serving any customer (Luo et al. 2015, Braekers and Kovacs 2016).

Regional consistency. In practice, and especially in urban contexts, the efficiency of a route depends on the driver’s familiarity with the addresses and buildings in the area, the typical traffic conditions on important streets or junctions, possible shortcuts or detours, etc. (Holland et al. 2017). As a result, it is desirable to maintain some form of regional consistency so that drivers can become familiar with their assigned or most common service regions and hence benefit from the associated learning effects (Haughton 2002, Zhong et al. 2007). This can be seen as a generalization of person-oriented consistency, and the previously mentioned fixed
routes can also be a way to delineate fixed regions (Wong and Beasley 1984). Similarly, regional consistency can be enforced by constructing routes that maximize the number of visited nodes within some threshold distance to fixed master routes (Sungur et al. 2010), or maximize the number of times a driver repeatedly visits the same region (Smilowitz et al. 2013).

**Delivery consistency.** In contexts such as vendor-managed inventories (Day et al. 2009), it may be desirable to deliver a consistent quantity of materials or provide a consistent level of service. Since cost-minimizing solutions do not typically possess these properties, Coelho et al. (2012) propose to constrain delivery quantities within lower and upper bounds, or to follow an order-up-to policy.

**Inconsistency.** Finally, inconsistency can be desired in some applications. For cash-in-transit operations, the routes should be unpredictable from day to day to reduce the risk of robberies (Bozkaya et al. 2017, Constantino et al. 2017). For the transportation of hazardous materials, safe backup routes should be available in case of adverse weather conditions or to spread the accident risk geographically (Akgün et al. 2000). In the \$m\$-peripatetic VRP, complete dissimilarity is ensured by requiring alternative solutions to be edge-disjoint (Ngueveu et al. 2010a,b). The \$k\$-dissimilar VRP relaxes this constraint and minimizes the average ratio between the length of edges shared by any pair of alternative solutions and the length of the routes containing those edges (Talarico et al. 2015). Martí et al. (2009) proposed a vertex-based dissimilarity measure, maximizing for each pair of paths the average distance between the vertices in one path and their closest neighbor in the other. Zajac (2016) considers geographic dissimilarity explicitly by minimizing the intersection of the geographic units visited in different solutions. Michallet et al. (2014), Hoogeboom and Dullaert (2019) and Soriano et al. (2019) enforce temporal inconsistency by using time-window penalties when the arrival times of consecutive visits at the same customer do not differ by more than a given constant.

### 2.5 Simplicity

In complex real-life systems, the acceptance and efficient realization of vehicle routing plans often depends on their simplicity and their intuitive appeal (Poot et al. 2002). As non-experts in combinatorial optimization, drivers and dispatchers may be reluctant to trust and implement solutions that appear overly complex and counter-intuitive or that require a high degree of coordination, even if these solutions are technically optimal with respect to cost. In routing contexts, visual appeal is often synonymous with compact and non-overlapping tours (Hollis and Green 2012). This corresponds to the two fundamental notions of clustering in the machine learning literature (Jain 2010): group together elements in such a way that similar elements (nearby deliveries) belong to the same cluster (compactness) and different elements (distant deliveries) belong to different clusters (separation).
Compactness. A route whose customers are geographically clustered is intuitive, because its compactness serves as a visual surrogate for low cost (Rossit et al. 2019). This intuition is typically exploited by optimization algorithms within the cluster-first, route-second category. In numerical terms, compactness is usually optimized at the route level by minimizing a measure of geographic spread. For example, Poot et al. (2002) minimize the average pairwise distance between all customers in a route, or the average distance of the customers to the route’s center of gravity. A different definition of “route center” is proposed by Tang and Miller-Hooks (2006), who define the median of a tour to be the customer that minimizes the maximum distance to any other customer in the same tour. These ideas correspond to the more general $k$-means and $k$-median clustering approaches. Likewise, although route compactness measures concern the operational level, they are closely connected to tactical decisions related to the design of compact distribution territories such as those commonly used in postal deliveries (see Section 3.1). The corresponding compactness criteria are similar, e.g., minimizing the maximum distance of a customer to their territory’s (route’s) center (Ríos-Mercado and Fernández 2009), the maximum distance between any pair of customers in the same territory (Lin et al. 2017), or the ratio of the territory’s (route’s) perimeter to the total perimeter of the service area (Lei et al. 2015). Other criteria are based on geometric ratios to ideal shapes like squares and circles (Kalcsics 2015) or even temporal characteristics such as time-window differences (Schneider et al. 2015).

Separation. Routes that are geographically separate make coordination easier, because local changes (e.g., unexpected demand) do not impact the rest of the plan (Lum et al. 2017). Moreover, if the geographic separation is done at the tactical level, then processes such as sorting can be executed in parallel with routing, reducing delivery times and improving competitiveness (Janssens et al. 2015). Unlike compactness, separation measures are calculated at the solution level, as they concern the relationship between multiple routes. They all minimize some measure of overlap. Example metrics include the number of customers that are closer to another route’s center or median than to their own (Poot et al. 2002, Tang and Miller-Hooks 2006), the number of edges shared by two or more routes in an arc routing context (Constantino et al. 2015), the number of customers contained in the convex hull of a route that is not their own (Poot et al. 2002), the average overlap of the routes’ convex hulls (Lum et al. 2017), the number of times different routes cross paths (Poot et al. 2002), and others (Corberán et al. 2017). Although these metrics may initially appear somewhat ad hoc, they can have meaningful properties. For example, Tang and Miller-Hooks (2006) show that if no customer is closer to another route’s median than to its own, then the convex hulls of the routes cannot overlap.

Navigation complexity. Routes should be easy to follow and execute. Distribution companies such as UPS prefer simple route structures so that drivers spend less effort on spatial route cognition and instead concentrate on driving safely (Holland et al. 2017). Users of consumer navigation systems prefer routes that are concisely described and can be easily followed, especially when traveling through unfamiliar environments (Shao et al. 2014). In practice, metrics for quantifying the navigation complexity of a route are commonly based on the number
and type of turns encountered. Turn restrictions and turn penalties frequently arise in arc routing applications (Assad and Golden 1995, Benavent and Soler 1999, Corberán et al. 2002, Vidal 2017) and can be refined by considering different types of intersections as well as the road network hierarchy (Duckham and Kulik 2003).

## 2.6 Reliability

Deterministic VRPs consider that all problem information is available and accurate. However, data are always subject to approximations, and unexpected events can render “optimal” deterministic routing plans inefficient or impracticable. As a consequence, finding reliable routing solutions that remain effective in the presence of uncertainty has become a major concern (Gendreau et al. 2014, 2016). Under uncertainty, a natural but cost-ineffective strategy is to use a deterministic model to generate reliable solutions that contain some slack (e.g., capacity or time). A better option is to exploit additional knowledge of the uncertain events, in the form of representative scenarios, probability distributions, or uncertainty sets, giving rise to stochastic or robust VRP models. Beyond a mere choice of objective function, defining a stochastic VRP requires to specify when and how stochastic parameter values are observed, and when decisions are taken. Two main groups of approaches can be distinguished: 1) stochastic programming models, which typically focus on minimizing the expected cost of the routes and recourse actions made as a consequence of uncertain events; and 2) chance constraints or robust formulations, which impose constraints on the failure probabilities.

**Expected cost or loss.** Stochastic models based on a priori optimization (Bertsimas et al. 1990) assume that the routing decisions are made in a first stage based on partial knowledge of future events (before any stochastic parameters are observed), and that prespecified recourse policies will be used in a second stage when unexpected events occur (e.g., a direct return to the depot in the case of excess demand). Most models in this family focus on optimizing the expected cost of first-stage routes and second-stage recourse actions. There are three main sources of uncertainty: customer demands (Bertsimas 1992), service requests (Jaillet 1988), and travel times (Laporte et al. 1992). Yet, despite considerable algorithmic progress over the last four decades, the solution methods (metaheuristics and mathematical programming methods alike) are limited by the necessity to evaluate the expected cost of the recourse actions. Therefore, strong assumptions are typically made to keep the evaluations tractable: simplistic recourse policies are used, and the probability distributions associated with the random events are assumed to be independent. Ongoing research is exploring more sophisticated recourse policies (Yang et al. 2000, Ak and Erera 2007, Louveaux and Salazar-González 2018, Salavati-Khoshghalb et al. 2019), correlated random events (Rostami et al. 2017), and multiple decision stages (Dror et al. 1989, Goodson et al. 2013).

**Risk of failure.** Models based on chance constraints or robust optimization impose constraints on the probability of failure as opposed to optimizing the expected cost of the uncertain
events. These approaches significantly differ in how they model stochastic parameters. Chance constraints still rely on distributional information to evaluate and bound the probabilities of failure. This paradigm has been commonly used to solve VRPs with stochastic travel times and time windows (Laporte et al. 1992, Li et al. 2010). However, as highlighted in Errico et al. (2018), there is a thin line between model assumptions that allow for efficient calculations (e.g., convolutions and dominance properties) and those that lead to intractable problems. In contrast, robust models rely on an uncertainty set (e.g., a polytope) to represent reasonable parameter variations and seek solutions that are feasible for any parameter realization within this set (Ben-Tal et al. 2009, Bertsimas et al. 2011). Robust models are especially useful in situations where no complete distribution information is available, and they are typically easier to solve than their chance-constrained counterparts (Sungur et al. 2008, Gounaris et al. 2013, Pessoa et al. 2018b). Since these models are completely risk-averse, research continues on alternative models of uncertainty (e.g., distributionally robust models) that are meaningful in practice and remain tractable (Jaillet et al. 2016, Zhang et al. 2019).

2.7 Externalities

Although transportation is essential for modern businesses and society, it also has undesirable consequences (Demir et al. 2015). A more holistic optimization of logistics and mobility is needed to mitigate the impacts of externalities while maintaining efficient transportation systems.

Emissions. Road transportation is a major contributor to increasing atmospheric pollution caused by greenhouse gases and particulates. Reflecting also the broader societal concerns about sustainability, the past decade has seen a rapid growth in studies falling under the class of green VRPs that account for emissions in the optimization model (Demir et al. 2014a). It has indeed been recognized that classical cost-minimizing objectives (in terms of distance or time) do not lead to minimal emissions or fuel consumption, although there is a correlation (Bektaş and Laporte 2011). A variety of fuel consumption models and solution methods have therefore been put forward (Demir et al. 2011, 2014a, Kramer et al. 2015, Fukasawa et al. 2018). Due to the complexity of the emissions functions, optimization methods need to handle various factors, e.g., load-dependency (Kara et al. 2007), time-dependency (Jabali et al. 2012b, Franceschetti et al. 2013), heterogeneous fleets (Koç et al. 2015), and modal choice (Bauer et al. 2010). Although direct speed optimization is difficult to plan for road-based operations, it is an important concern and easier to achieve in maritime transportation (Fagerholt et al. 2009, Norstad et al. 2011, Hvattum et al. 2013). From a practical perspective, it is worth noting that by allowing a small increase in distance or time, one can significantly reduce emissions, which motivates the consideration of fuel consumption as a side objective (Demir et al. 2014b).

Safety Risks. When transporting hazardous materials (hazmat) such as nuclear waste, chemical agents, or noxious gases, risk mitigation is a priority. Since the degree of risk and the
severity of a potential accident are closely related to the selected route, classical VRP models must be carefully extended to properly incorporate various aspects of risk. For example, Tarantilis and Kiranoudis (2001) consider population exposure risk on each link of the network, Ma et al. (2012) propose the inclusion of link-specific risk capacities, and Taslimi et al. (2017) examine a bilevel problem in which a regulator decides which links to close for hazmat transportation while considering the expected alternative routes then chosen by the hazmat carriers. Accident risk is also considered along the temporal dimension by Meng et al. (2005) and Toumazis and Kwon (2013), who consider time-dependent risk models, and by Zografos (2004), who examines the trade-off between travel time and risk. Note that some hazmat VRP models can be generalized to different types of undesirable externalities, such as noise, disturbance, and pedestrian safety (Bronfman et al. 2015, Grabenschweiger et al. 2018). Finally, consumer-oriented routing applications optimize safety from the opposite perspective, aiming to generate routes that are safe for the user (Shah et al. 2011, Kim et al. 2014).

3 Integrated Problems – Routing as an Evaluation Tool

Vehicle routing decisions are fundamentally operational but are often linked with other decisions taken at a strategic or tactical level over a longer planning horizon (Crainic 2002). In such contexts, generating VRP solutions or at least evaluating their characteristics becomes essential to evaluate the cost of planning decisions made at a higher level, which can be districting, facility location, fleet composition, or inventory and production management. Two main approaches are typically used: continuous approximation and regression models (Franceschetti et al. 2017b), or fast versions of VRP algorithms adapted to stochastic or scenario-based problem variants. The former aims to give a good estimate of the routing costs based on geometric considerations, while the latter samples demand patterns resulting from distribution or scenario information. While stochastic and scenario-based approaches offer greater precision, they generally lead to large-scale integrated problems which challenge the capabilities of current solution methods. This section surveys the main applications and methods arising in integrated two-level problems of which routing is one of the components.

3.1 Routing and Districting

Districting is the process of partitioning a territory for political, administrative or commercial purposes (Kalcsics 2015). The best-known application is the design of political districts (see, e.g., Bozkaya et al. 2003), but logistics applications are also common. These include the design of sales territories (Skiera and Albers 1998, Drexel and Haase 1999, Lei et al. 2015) and distribution management applications, for example those encountered in mail delivery systems (Rosenfield et al. 1992, Novaes et al. 2000, Bruno et al. 2019). Fixed districts ensure regional consistency and facilitate delivery operations (see Section 2.4). Districting plans are typically subject to hard constraints such as contiguity as well as soft constraints such as size, compactness, population balance, homogeneity, and fairness. These soft constraints, which are often
nonlinear, are eventually aggregated into a multi-criteria objective function with suitable user-defined weights. The districts are often expected to change over time because of population shifts, for example. In such cases, robustness with respect to future stochastic or dynamic changes is also deemed to be a desirable property (Lei et al. 2016).

There are two main techniques for constructing districts. The most common aggregates cells, usually called basic units, for which geographic, demographic or socio-economic data are available. It is common to define basic units as census tracts (Bozkaya et al. 2003). A second technique divides a planar area by drawing lines that define the district boundaries through geometric arguments, as in Carlsson (2012) and Carlsson and Delage (2013). One obvious advantage of the first technique is that it lends itself to the use of local search-based metaheuristics in which basic units are iteratively relocated or swapped between adjacent districts and allows efficient evaluations of the objective function.

We focus on applications in which a traveling salesman problem (TSP) or VRP must eventually be solved within each district. A common case arises in the planning of sales districts, where each district is assigned to a vendor or a team of vendors. When designing the districts, one must take into account the routing cost and also ensure a level of equity between the routes of different districts. If, as is usually the case, a local search technique is used to optimize the districts, it can be prohibitively long to optimize the vehicle routes associated with the districts at each step (i.e., move evaluation). To circumvent this issue, most solution methods rely on closed-form formulas to approximate the routing costs without actually determining the routes. Two such formulas are the Beardwood-Halton-Hammersley (BHH) formula for the TSP (Beardwood et al. 1959) and the Daganzo (1984) formula for the VRP. The BHH formula approximates the routing cost through \( n \) independently and identically distributed points in a compact area of size \( A \) as \( \beta \sqrt{nA} \), where \( \beta \) is a constant. Appropriate constant values are provided in Applegate et al. (2011). Combining this formula with a simple geometrical partitioning strategy, Daganzo (1984) approximates the cost of a VRP solution as \( 2rm + 0.57 \sqrt{nA} \), where \( m \) is the number of vehicle tours and \( r \) is the average distance between the depot and the barycenters of the districts. The first term in this expression represents the “line-haul” distance to reach the districts, and the second term measures the routing costs within the districts. Continuous approximation formulas are still being refined and generalized (see, e.g., Çavdar and Sokol 2015, Merchán and Winkenbach 2019), and approaches based on regression or neural networks (see, e.g., Kwon et al. 1995) may soon achieve even better trade-offs between estimation accuracy and computational effort.

### 3.2 Routing and Facility Location

Many applications require the evaluation of routing costs during the facility location decisions. This has led to the development of a vast literature dedicated to combined location and routing problems (Prodhon and Prins 2014, Laporte et al. 2015, Schneider and Drexel 2017). Facility location decisions are strategic or tactical in most applications. They concern warehouses, cross-docks, or satellite facilities in city logistics, whereas vehicle routes are operational de-
cisions that can change dynamically over time. In these contexts, continuous approximation formulas can be used to estimate the routing cost (Laporte and Dejax 1989, Campbell 1990, Ouyang and Daganzo 2006, Xie and Ouyang 2015), and facility catchment areas may be represented as polygons in a Voronoi diagram (Laporte and Dejax 1989). Continuous approximation methods can also be extended to integrate a variety of constraints and objectives, such as backbone costs in hub networks (Campbell 2013, Carlsson and Jia 2013, 2015).

Another approach for location and routing is to rely on Monte Carlo scenario generation as a basis for routing cost evaluations (Klibi et al. 2010). This approach, however, can lead to challenging scenario generation and optimization problems. This may explain why most studies on combined location and routing problems have opted for a deterministic “single routing scenario” approach, giving rise to the canonical location-routing problem (LRP), recently surveyed in Schneider and Drexl (2017). The LRP model is mainly relevant in contexts where the delivery routes are fixed over a long time, or where both location and routing decisions are operational, e.g., when locating transfer points between two vehicles or vehicle reception points such as temporary parking places and postal boxes (Boudoin et al. 2014). The canonical LRP represents a challenge for exact algorithms (Baldacci et al. 2011, Contardo et al. 2014), since these approaches must ultimately enumerate many candidate subsets of locations. In contrast, metaheuristics currently produce good solutions for large-scale instances (Schneider and Löffler 2019). In the future, these methods may be extended to sophisticated settings with multiple routing scenarios in an attempt to improve the accuracy and applicability of tactical location routing models.

3.3 Routing and Fleet Composition

Tactical fleet sizing and composition problems occur across all transportation modalities, when renewing vehicles, adapting to market fluctuations, and evaluating business changes (e.g., company mergers). Fleet size adjustments can be done via long-term vehicle acquisitions and sales or short-term leasing. Typical planning horizons vary among applications: horizons are generally longer in maritime operations than in land-based transportation because of the long lifetime of ships and the large capital costs incurred (Hoff et al. 2010). As a result, maritime fleet sizing models usually consider fixed trade lanes for strategic planning (Pantuso et al. 2014, Wang et al. 2018). For land-based transportation, two main approaches are generally used to evaluate the routing costs within fleet composition models: continuous approximations or (multi-period or stochastic) heterogeneous VRP solution methods.

Continuous approximation models stem from the observation that it is difficult to obtain accurate demand scenarios and even harder to solve the resulting VRPs. Time-consuming route evaluations can therefore be avoided by the use of approximation formulas to focus the optimization on the fleet sizing decisions (Campbell 1995, Jabali et al. 2012a, Franceschetti et al. 2017a, Nourinejad and Roorda 2017).

Heterogeneous VRP models, in contrast, require the joint determination of vehicle types and routes. Each vehicle type may possess distinct characteristics, e.g., capacity, fixed and variable
costs, customer-service restrictions, or even specific travel costs and speeds. Two canonical problems are generally distinguished: the fleet size and mix VRP (FMVRP) and the heterogeneous fixed fleet VRP (HFVRP). The FMVRP assumes that an unlimited number of vehicles of each type is available, whereas maximum limits are set in the HFVRP. As illustrated in the survey of Koç et al. (2016), research on heterogeneous VRPs is extensive but usually focused on a single period in the presence of a fixed set of customer requests. This case corresponds to applications in which the fleet is already acquired (or rented for a short term) or where the demand is stable over a long time period. Kilby and Urli (2016), Pasha et al. (2016), and Bertoli et al. (2019) have recently extended the FMVRP to multi-period and stochastic settings, helping to bridge the gap between the heterogeneous VRP and its tactical fleet composition applications.

Finally, the emergence of vehicles with alternative fuels and the growing focus on (locally) emission free deliveries have led to new fleet composition problems involving battery-powered and conventional vehicles (Felipe et al. 2014, Pelletier et al. 2016, Hiermann et al. 2016). Cities around the world are gradually restricting the vehicle types allowed in city centers. To cope with these challenges, there have been studies of fleet composition models with city center restrictions (Davis and Figliozzi 2013, Franceschetti et al. 2017a, Hiermann et al. 2019a). Another transition is taking place between transporter-managed and crowdsourced delivery systems. Crowdsourcing involves paying daily commuters and ad hoc drivers for last-mile deliveries in an effort to use their residual capacity, leading to a new generation of tactical fleet composition and multi-modal transportation problems (Archetti et al. 2016, Arslan et al. 2019, Cleophas et al. 2019, Mourad et al. 2019).

### 3.4 Routing, Inventory, and Production Management

Inventory-routing problems (IRPs) arise in the context of vendor-managed inventory management in which a supplier jointly optimizes vehicle routes, delivery schedules, and quantities. The field is rooted in the work of Bell et al. (1983) and has since seen a phenomenal growth, discussed in the survey of Coelho et al. (2014). Multiple versions of the problem exist, varying in the planning horizon (finite or infinite), the delivery structure (1-1, 1-M, M-M, or 1-M-M-1: see Section 4.3), the routing patterns (back and forth routes or multi-customer routes), the inventory policy (maximum level or up-to-order), the inventory decisions (lost sales or backlogging), the fleet composition (homogeneous or heterogeneous), and the fleet size (single, multiple, or unconstrained). Since the planning horizons tend to be shorter in IRPs than in other strategic problems discussed in this section, a larger part of the literature combines inventory management with route generation within integrated VRP models, although continuous routing-cost approximations are also sometimes used (Baller et al. 2019). Most models are defined on a rolling horizon, so the choice of objective is nontrivial. In particular, optimizing the logistic ratio (Archetti et al. 2017b) can be better in practice than pure cost minimization. IRPs are notoriously challenging for exact methods (Desaulniers et al. 2016), but there are efficient hybrid metaheuristics (see, e.g., Archetti et al. 2017a, Chitsaz et al. 2019).

As discussed in the surveys of Christiansen et al. (2013) and Papageorgiou et al. (2014),
IRPs have often been applied to maritime routing, particularly for the transportation of liquefied natural gas (see, e.g., Stålhane et al. 2012, Halvorsen-Weare and Fagerholt 2013, Andersson et al. 2016, Ghiami et al. 2019). Another important application is the transportation of perishable products (see, e.g., Coelho and Laporte 2014, Crama et al. 2018). Moreover, recent years have seen the emergence of inventory-routing problems related to the management of shared mobility systems, mostly in the case of bikes and cars. In these challenging problems, one must simultaneously optimize the inventory levels at the stations and the itineraries used to reposition the shared vehicles (Chemla et al. 2013, Laporte et al. 2018).

Finally, supply chain integration extends well beyond inventory routing. As demonstrated by Chandra (1993) and Chandra and Fisher (1994), the joint optimization of routing, inventory management, and production can lead to substantial savings over a sequential approach. The resulting production-routing problem (PRP) aims to coordinate a production schedule with product deliveries at customer locations (Adulyasak et al. 2015). Recent algorithms and case studies are presented in Adulyasak et al. (2014), Absi et al. (2015), Neves-Moreira et al. (2019) and Qiu et al. (2019). Ongoing research is considering integrating a wider set of supply chain decisions, e.g., assembly, production, inventory, and routing (Chitsaz et al. 2019), or production, location, and inventory (Darvish and Coelho 2018).

4 Refined Problems – Precise and Applicable Plans

In parallel with studies that concern the integration of VRP models with other tactical supply chain decisions, significant research is being conducted to refine the models and integrate fine-grained problem attributes that can have a large impact on solution quality and feasibility. This section reviews some important problem refinements in relation to the transportation network, the drivers and vehicles, and the customer requests.

4.1 Specificities of the Transportation Network

Arc attributes. Transportation networks are usually characterized by multiple attributes, including driving time, driving cost (and tolls), transportation mode, attractiveness, safety, emissions, and energy consumption. In these conditions, a single best path may not be readily definable between each origin and destination, and several trade-off paths should be considered. For example, the canonical VRP with time windows has been extensively studied with the fundamental assumption that one time unit corresponds to one cost unit. In such situations without any trade-off, the search can be limited to a single shortest path for each origin and destination. However, time and cost are not directly proportional in real transportation networks: these resources can even be negatively correlated when tolls or access restrictions are imposed (Reinhardt et al. 2016). Research on this topic is fairly recent. Accounting for these effects gives rise to a class of VRPs on multi-graphs (Ben Ticha et al. 2017, 2018, 2019, Hiermann et al. 2019a, Soriano et al. 2019) linked to critical applications in multi-modal transportation, long-haul transportation, and city logistics, among others (Caramia and Guerriero
Solution methods must jointly optimize the visit sequences and the paths between them. In the worst case, the number of trade-off paths between any two points grows exponentially. Still, empirical analyses have shown that this number remains small in practice for transportation networks with time-window constraints (Müller-Hannemann and Weihe 2006, Ben Ticha et al. 2017), and the set of paths could otherwise be heuristically restricted (Hiermann et al. 2019a).

Two-echelon structures. Studies on distribution networks possessing a two-echelon structure can be traced back to the work of Jacobsen and Madsen (1980), in which intermediate facilities are used to transfer newspapers from large vehicles to smaller ones. Nowadays, as reviewed in Cuda et al. (2015) and Guastaroba et al. (2016), e-retailers commonly adopt a two-echelon structure to deliver orders from distribution centers to cross-docking facilities for consolidation, and thence to customers. In city logistics, transfer points are typically located on the outskirts of cities to reduce noise, pollution and traffic (Soysal et al. 2015). Research on two-echelon VRPs is now very active since the joint optimization of two route levels and the related time constraints and synchronization issues pose substantial methodological challenges (Grangier et al. 2016). We refer to Breunig et al. (2016) and Marques et al. (2019) for state-of-the-art heuristic and exact algorithms.

Congestion and time dependency. Congestion is a major factor in city logistics, since it causes massive economic losses (400 billion dollars per year in the United States according to Cookson and Pishue 2018) and has numerous negative effects. As noted in the survey of Gendreau et al. (2015), VRPs with time-dependent travel times (TDVRPs) may arise as a consequence of congestion, weather conditions, road closures, roaming targets, and other factors. TDVRPs have been the focus of extensive research, but the recent survey of Rincon-Garcia et al. (2018) reports that the inadequate management of time-dependent travel times in routing software remains a major barrier to application. Time-dependent effects are commonly modeled via travel-time or travel-speed functions (Gendreau et al. 2015). Furthermore, the speed on an arc may be computed at its entry time (frozen link model) or may vary on the arc as time passes (elastic link model). In an elastic link model with strictly positive speeds, the FIFO property is always satisfied, i.e., a later departure leads to a later arrival time (Orda and Rom 1990). This model has been used in the seminal work of Ichoua et al. (2003).

Most studies on TDVRPs rely on a complete graph representation of the network in which each origin-destination pair is represented by a single link and travel time function. In practice, however, the time-dependent travel times are specific to each street or neighborhood of an urban network. To account for this, some studies have defined time-dependent speed functions at the network level (Maden et al. 2009, Huang et al. 2017, Vidal et al. 2019). It is important to note that most existing vehicle routing heuristics can be adapted to the TDVRP under the condition that a fast mechanism is available for time-dependent travel time queries. Yet, despite the development of sophisticated quickest path algorithms (Batz et al. 2013, Bast et al. 2016), producing accurate speed predictions and performing rapid travel-time queries on large-
scale networks (typically within a fraction of a millisecond) raise significant methodological challenges.

**Access restrictions.** Turn restrictions, delays at intersections, tolls, and limited parking availability are a significant part of the reality of urban logistics. The inadequate management of these aspects is another important barrier to the application of routing software in practice (Rincon-Garcia et al. 2018). Nielsen et al. (1998) estimate that turns and delays at intersections represent 30% of the total transit time in cities, so an accurate model of turn restrictions is critical for mail delivery, waste collection, snow plowing, and street maintenance operations, among others (Perrier et al. 2008, Irnich 2008). Likewise, an excessive number of turns in warehouse operations can lead to increased chances of vehicle tipovers, congestion, and collisions (Çelik and Süral 2016).

Accounting for these detailed effects is not straightforward. In the case of turn restrictions, for example, joining turn-feasible shortest paths may still lead to forbidden turns at their junctions. Solution approaches for such problem variants rely on graph transformations (Clossey et al. 2001, Corberán et al. 2002, Vanhove and Fack 2012) or exploit a *mode selection* subproblem to optimize the arrival direction at each service location during route evaluations (Vidal 2017). Exact algorithms may require dedicated pricing and cut separation procedures to consider costs based on consecutive edge pairs (Martinelli and Contardo 2015). The limited amount of space in city centers also leads truck drivers to rely on double parking. Some recent studies have modeled the impact of such practices (Morillo and Campos 2014, Figliozzi and Tipagornwong 2017), yet parking considerations remain largely unrepresented in VRP models.

### 4.2 Specificities of Drivers and Vehicles

**Heterogeneous vehicles and delivery modes.** As discussed in Section 3.3, vehicle fleets are rarely homogeneous (Pantuso et al. 2014, Koç et al. 2016). Individual vehicle specificities (e.g., variable costs, specific equipments, or access restrictions) must often be explicitly considered to obtain accurate operational plans, giving rise to FMVRP and HFVRP variants. Many efficient metaheuristics and exact algorithms have been proposed for these problems (see, e.g., Vidal et al. 2014, Koç et al. 2015, Pessoa et al. 2018a, Penna et al. 2019). Recent studies have extended the scope of heterogeneous VRPs to multi-modal transportation systems involving bikes, scooters, vans, as well as alternative propulsion modes (Felipe et al. 2014, Nocerino et al. 2016, Hiermann et al. 2019b). Beyond this, the recent growth of e-commerce has given rise to new distribution practices, including the use of drones. In the simplest case, drones make back-and-forth deliveries from a warehouse to customer locations. More sophisticated distribution modes involve the combined use of delivery vehicles and drones. For example, Murray and Chu (2015), Dorling et al. (2017), Poikonen et al. (2017), and Agatz et al. (2018) consider a delivery configuration in which a drone, mounted on a vehicle, detaches itself to perform deliveries while the vehicle keeps moving. The resulting problems are gradually giving rise to a rich research
Working hours regulations. Hours-of-service (HOS) regulations are ubiquitous, and they should be taken into account when long-haul routes are generated for several days or weeks. Transportation companies, in particular, have the responsibility of ensuring that driving plans can be safely performed with regulatory break and rest periods. Typical HOS regulations in the United States, the European Union, Canada, and Australia impose daily and weekly rest periods as well as limits on the driving and working hours. Their numerous clauses, conditions, and exclusions make it extremely difficult to check that a compliant schedule exists, even for a fixed sequence of visits. Prescott-Gagnon et al. (2010) and Goel and Vidal (2014) have studied these rule sets for different countries and proposed efficient routing and scheduling algorithms. The latter study, in particular, used the optimized routing plans to compare various regulations in terms of their impact on drivers’ fatigue.

HOS regulations also extend beyond classical single-driver day operations, and specific provisions exist for night work (Goel 2018) and team-driving (Goel et al. 2019). Schiffer et al. (2017) recently highlighted the benefits of jointly planning rest periods and recharging actions for electric vehicles. A key challenge of HOS regulations relates to the purposeful use of optimization: transportation companies should verify that a feasible schedule exists, but most decisions on break and rest periods lie with the drivers. In such situations, a simple simulation of driver behavior may be more reliable than a full-blown optimization algorithm considering all regulatory aspects and exceptions. Beyond this, there is a thin line between regulatory aspects that can be optimized and those that should be used as a recourse when facing unforeseen events (e.g., the extended driving time defined by regulation (EC) 561/2006 should likely be kept as a recourse).

Loading constraints and compartments. Trucks, ships, and airplanes have many specific load restrictions which must be taken into account during optimization (Pollaris et al. 2015). The papers considering these aspects are primarily classified by geometry, e.g., pallet loading (Pollaris et al. 2017), 2D packing (Iori et al. 2007), and 3D packing constraints (Gendreau et al. 2006), but other constraints related to fragility, orientation, or equilibrium often come into play. Specialized applications such as car hauling require dedicated feasibility-checking mechanisms to ensure that a load can be feasibly placed on the truck (Dell’Amico et al. 2015) and that axle-weight limits are respected (Pollaris et al. 2017). Many of these VRP variants share the common trait that load-feasibility checking, even for a fixed route, is an NP-hard problem. To speed up this critical evaluation step, a variety of packing heuristics, bounds, and rules may be used to directly filter some feasible or infeasible loads. Moreover, the loading constraints go well beyond the search for a feasible packing of items: some applications require precedence constraints (e.g., LIFO or FIFO) between services to make unloading possible (Cordeau et al. 2009) or integrate handling constraints for on-board load rearrangement (Battarra et al. 2010), while other applications, e.g., for hazardous materials or food transportation, impose incompatibility or separation constraints (Battarra et al. 2009, Hamdi-Dhaoui et al. 2014). The loading area can also be unique, split into different compartments (Derigs et al. 2010), or even sepa-
rated into a truck and a trailer (Villegas et al. 2013). The trailer can be parked and retrieved to facilitate access to some customers, leading to two-echelon problem variants.

**Recharging stops.** There has been a rapid growth of research into VRP variants for battery-powered electric vehicles. Because of their limited range, early electric models often required en route recharging stops, and these intermediate stops (Schiffer et al. 2019) became a defining feature of most electric VRPs (EVRP). The EVRP literature has quickly grown to take into account the numerous characteristics of real applications. Studies have been conducted on EVRPs with heterogeneous fleets and charging infrastructure (Felipe et al. 2014, Hiermann et al. 2016, 2019b); more realistic energy consumption functions (Goeke and Schneider 2015) and charging profiles (Keskin and Çatay 2016, Montoya et al. 2017); limited charging capacity (Froger et al. 2017); and time-dependent energy costs (Pelletier et al. 2018). As battery technology progresses, the range of electric vehicles is becoming sufficient for daily delivery operations in metropolitan areas. Therefore, en route recharging is gradually disappearing from these applications. It may still be necessary for lightweight vehicles (e.g., drones) or vehicles performing round-the-clock operations. Also note that the limited supply of some materials (e.g., rare earths) and the lack of a good recycling process can limit the availability of large batteries (Hwang et al. 2017), so the development of a more efficient recharging infrastructure remains a plausible scenario.

### 4.3 Specificities of Customer Requests

Some customer-request specificities arising in the form of customer-oriented objectives have been discussed in Section 2. Here we discuss other aspects of customer requests that do not arise as an optimization goal but are nevertheless essential for useful routing plans.

**Service types.** VRP applications can involve very different service types, depending on the number of commodities involved and on the origin and destination points (depot or customer location). Four main types can generally be distinguished:

- **1-M-1 (including 1-M and M-1).** One-to-many-to-one problems include depot-to-customer and customer-to-depot transportation as special cases. Applications of 1-M-1 services arise, e.g., in small-package delivery, where deliveries are made early in the routes and are followed by pickups later in the day (Holland et al. 2017).

- **1-1.** One-to-one problems represent transportation settings in which each service is unique and associated with a fixed origin and destination. A typical application is taxi fleet operations (Doerner and Salazar-González 2014).

- **M-M.** Many-to-many problems involve one or several resources at multiple locations. The goal is to move some of these resources toward the locations where they are most needed. Typical applications concern bike repositioning (Chemla et al. 2013, Bulhões et al. 2018b) and lateral transshipments (Paterson et al. 2011, Hartl and Romauch 2016).
• **1-M-M-1.** Finally, some applications may involve a combination of the M-M and 1-M-1 cases. One such problem was investigated by van Anholt et al. (2016) in the context of the replenishment of automated teller machines: money has to be transferred from a central office to automated tellers, among these tellers, and back to the office. As noted in Battarra et al. (2014), the first two categories of problems (1-M-1 and 1-1) have been extensively discussed in the VRP literature. In contrast, studies on M-M or 1-M-M-1 settings, especially with multiple commodities, are not as common.

Applications also differ in terms of whether or not split shipments are allowed. Split loads (Archetti and Speranza 2012) typically occur when transporting many units of the same commodity (e.g., bikes) or when delivering or collecting a divisible product (e.g., food or liquids). In such situations, a customer may be visited multiple times in order to fulfill its request. The resulting split delivery VRP is a relaxation of the capacitated VRP (CVRP) but is more complicated to solve. Because of a lack of an efficient route-based decomposition, due to the customer demands which act as linking constraints, exact methods (Archetti et al. 2014a) struggle to optimally solve instances of a size (e.g., 50 to 100 deliveries) that can easily be handled in a canonical CVRP setting. Finally, applications involving pickups and deliveries with split loads have been considered in Nowak et al. (2008), Sahin et al. (2013) and Haddad et al. (2018). This setting is unexpectedly challenging: it is possible to create a family of benchmark instances for which any optimal solution requires a number of split pickups and deliveries that is an exponential function of the instance size (Haddad et al. 2018).

**Time constraints.** A wide range of VRP variants arising from the addition of time restrictions were surveyed in Vidal et al. (2015). Time constraints can arise as customer- or self-imposed time windows for deliveries (Solomon 1987, Agatz et al. 2011, Jabali et al. 2015, Bruck et al. 2018). In addition, release and due dates for commodities can be imposed at the depot (Cattaruzza et al. 2016, Shelbourne et al. 2017). Both of these settings can be viewed as time-window constraints on pickup or delivery locations. Other applications impose response-time limits between a request and its fulfillment by a vehicle. This is critical for customer satisfaction in mobility-on-demand systems, or for the delivery of perishable products (Pillac et al. 2013). Finally, in dial-a-ride transportation services where passengers with distinct pickup and delivery locations share the same vehicle, ride-time constraints are typically imposed to limit detours for each customer (Cordeau and Laporte 2007, Paquette et al. 2012). These constraints, however, make feasibility checks more complex. Research is ongoing into efficient solution evaluation procedures for these problems, to speed up heuristic search using preprocessing, incremental evaluations, and concatenations (Tang et al. 2010, Vidal et al. 2015, Gschwind and Drexl 2019).

**Skills.** Finally, maintenance or home care services may require specific skills. These requirements must be taken into account when assigning technicians and vehicles to tasks (Cappanera et al. 2013, Paraskevopoulos et al. 2017, Xie et al. 2017). In complex situations, several vehicles and workers with different skills and equipment (Eveborn et al. 2009, Parragh and Doerner 2018) may be jointly requested for a single task, leading to VRPs with synchronization constraints;
these are notoriously difficult to solve (Drexl 2012). Synchronization may even be imposed between different technicians at different locations, as in an application to electric network recovery studied by Goel and Meisel (2013). In this setting, a local change in one route can impact the entire daily schedule, violating the synchronization constraints or delaying other routes.

5 Challenges and Prospects

As we have shown, extensive research has been conducted over 60 years to better connect vehicle routing models and application cases. This close proximity between academic research, software companies, and transportation actors has led to a multitude of successful applications (see, e.g., Toth and Vigo 2014, Hall and Partyka 2018). Nonetheless, vehicle routing research is far from a closed topic. Technologies and business models evolve at a rapid pace. The continuing growth of e-commerce and home deliveries, increased access to on-demand transportation via mobility applications, and ongoing urbanization have put city transportation networks and supply chains under an unprecedented strain. To meet these challenges, companies and governing authorities seek true shifts of transportation paradigms rather than incremental optimizations of existing systems. These changes may be linked to new transportation modes, e.g., drones (Poikonen et al. 2017) or autonomous vehicles (Fagnant and Kockelman 2015), or to the redesign of business models and supply chains, e.g., crowdsourced deliveries (Arslan et al. 2019) or the physical Internet (Montreuil 2011). Regardless of the technology adopted, whereas products, drivers, and customers were typically aggregated into a route in classical VRPs, future applications will increasingly differentiate, synchronize, and optimize multiple flows associated with products, customers, and vehicles. The efficient coordination of such systems is a challenging task, and the associated rise in complexity rests on a fine equilibrium: while optimization models and their data requirements should be as sophisticated as required, they should also remain as simple as possible.

With respect to methodology, the development of heuristics and mathematical programming algorithms that are simple and efficient yet general enough to cope with a wide gamut of VRPs remains a crucial topic. Significant progress has been achieved by disciplined research built on problem-structure analysis and decision-set decompositions (see, e.g., Vidal et al. 2014, Vidal 2017, Toffolo et al. 2019, Pessoa et al. 2019). There is also a need to scale up VRP research. Current algorithms are usually evaluated on benchmark instances with a few hundred delivery points. This size could be strategically increased to thousands of visits to reflect emerging applications (Uchoa et al. 2017, Arnold et al. 2019). Multiple planning periods and scenarios should also be considered when relevant, e.g., for districting or location-routing.

Finally, it is important to focus our energy on problem variants that are truly of methodological and practical interest. Indeed, solving a new VRP variant made up of an arbitrary combination of attributes is certainly a technical achievement, but it does not necessarily constitute a significant methodological advance. Reproducibility and benchmarking are other important concerns. Methodological issues such as over-tuning, as well as differences in coding protocols
and in hardware have been raised by several researchers, but are not yet fully resolved. The fact that some flagship journals now require that codes be submitted as a condition for paper acceptance should, in all likelihood, foster the enforcement of stricter experimental standards.

Acknowledgments

This research was partly funded by the Canadian Natural and Engineering Research Council [grant number 2015-06189] as well as the National Council for Scientific and Technological Development [grant number 308498/2015-1], CAPES and FAPERJ [grant number E-26/203.310/2016] in Brazil. This support is gratefully acknowledged. Thanks are also due to the referees for their valuable comments.

References


