WORKING PAPER SERIES

Optimal Service Time Windows

Marlin W. Ulmer/Justin C. Goodson/Barrett W. Thomas

Working Paper No. 1/2023

FACULTY OF ECONOMICS
AND MANAGEMENT

Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Der Dekan

Verantwortlich für diese Ausgabe:
M. W. Ulmer, J. C. Goodson and B. W. Thomas
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Postfach 4120
39016 Magdeburg
Germany

http://www.fww.ovgu.de/femm

Bezug über den Herausgeber
ISSN 1615-4274

Optimal Service Time Windows

Marlin W. Ulmer
Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany, marlin.ulmer@ovgu.de

Justin C. Goodson
Department of Operations & Information Technology Management
Saint Louis University, St. Louis, MO,

Barrett W. Thomas
Department of Business Analytics,
University of Iowa, Iowa City, IA

Because customers must usually arrange their schedules to be present for home services, they desire an accurate estimate of when the service will take place. However, even when firms quote large service time windows, they are often missed, leading to customer dissatisfaction. Wide time windows and frequent failures occur because time windows must be communicated to customers in the face of several uncertainties: future customer requests are unknown, final service plans are not yet determined, and when fulfillment is outsourced to a third party, the firm has limited control over routing procedures. Even when routing is performed in-house, time windows typically do not receive explicit consideration. In this paper, we show how companies can communicate reliable and narrow time windows to customers in the face of arrival time uncertainty. Under mild assumptions, our main result characterizes the optimal policy, identifying structure that reduces a high-dimensional stochastic non-linear optimization problem to a root-finding problem in one dimension. The result inspires a practice-ready heuristic for the more general case. Relative to the industry standard of communicating uniform time windows to all customers, and to other policies applied in practice, our method of quoting customer-specific time windows yields a substantial increase in customer convenience without sacrificing reliability of service, providing results that nearly achieve the lower bound on the optimal solution. Our results show that (i) time windows should be tailored to individual customers, (ii) time window sizes should be proportional to the service level, (iii) larger time windows should be assigned to earlier requests and smaller time windows to later requests, (iv) larger time windows should be assigned to customers further from the depot of operation and smaller time windows to closer customers, and (v) two time windows for one customer are helpful in some cases.

Key words: time windows, service routing, non-linear optimization

1. Introduction

While one of the authors writes this introduction, he sits at home close to the door waiting for a shipment to arrive. The delivery company gave him a time window from 3pm to 7pm, but he waits and writes at 9am because experience tells him that even though the communicated time window is already very large, there is no guarantee that it will be satisfied, and he does not want to miss the shipment (again). Waiting is not new for the authors. Recently, Marlin Ulmer scheduled internet installation with Deutsche Telekom. They offered service during a five-hour time window from 8am to 1pm. The wide time window required that he stay at home half the day so that he could receive the service technician. When Justin Goodson purchased an appliance, he was given a four-hour delivery time window from noon to 4pm. On the delivery date, Goodson received a phone call at 7am indicating arrival in 20 minutes. If he could not meet the driver, then delivery would have to be rescheduled for a different day. Barrett Thomas paid a premium for FedEx express package delivery. When delivery was attempted outside of the time window, Thomas was not present and the package was not delivered. After two more attempted deliveries, both outside of communicated time windows, FedEx returned the package to the vendor.

Many readers will share the authors' frustrations. Not only do providers of delivery, installation, and in-home repair services typically communicate very large time windows at the time of booking, they often fail to meet those time windows on the day of service. The reasons for wide time windows and frequent failures are twofold. First, the time window is communicated to a customer when they request service. Because requests arrive sequentially, the firm does not know what other customers will be scheduled for service on the same day as the given customer. Consequently, the schedule is set after time windows are communicated. Even then, uncertain travel and service times make it difficult to predict arrival times to customers. Second, because time is money, firms have invested years and millions to develop effective scheduling and routing algorithms. In competitive markets, cost-efficient routing solutions are essential (for example, see Prisco (2017)). Notably, the satisfaction of communicated time windows usually does not play a role in the routing optimization. Alternatively, whenever delivery operations are outsourced to third-party providers, the firm that originates the product or service typically has little say in route design. This helps to explain the practice of communicating very large time windows, as well as their frequent violations (Apte et al. 2007).

While cost-efficient routes save companies money, others pay. People receiving home services have to plan in advance and often take time off work to be home while service occurs. The real-time tracking some companies offer on the day of service does not help much in such cases because adjustments to work and child care schedules, for example, typically require more advance notification. One estimate puts the cost of people waiting for home services at $38-billion in 2011 (Ellis 2011). With the rise of e-commerce, this number is likely even higher today. Despite their use of routing and scheduling methods that disregard customers' desires for smaller and more reliable time windows, firms stand to benefit from movement in this direction. Missed time windows and customers require costly recourse actions (Lim et al. to appear). Reducing these incidents can benefit both providers and customers.

How can we change the current state of affairs? How can firms keep their carefully developed routing algorithms, or outsource logistics to third parties, but at the same time communicate shorter and more reliable time windows to customers? Answering this question is the focus of this paper. We consider the situation in which a firm providing on-location services schedules customers during a booking period and then executes the service at a later date. A time window is communicated when a customer requests service. Because customers randomly request service across the booking period, time windows must be determined under incomplete information about the set of customers that will eventually be served and, consequently, under incomplete information about the order in which the customers will be served. Thus, at the moment the time window is assigned, we know only a probability distribution on the arrival time to the customer. We make no assumption on how customers are routed. Rather, the arrival time distribution accounts for a firm's chosen routing method. For every customer request during the booking period, we set a time window for future service. Because the decision comprises a start and an end time, we also control the size of the time window. The goal is to minimize the expected time window size across all customers. Since providers draw on their existing scheduling and routing mechanisms or outsource service to third parties, in practice, time window decisions are decoupled from the final routing. Thus, we cannot guarantee that a time window communicated during the booking period will be satisfied for every customer. To ensure a measure of reliability, we follow practice and impose a chance constraint to ensure that a high percentage of time windows are satisfied (e.g. for 95 percent of the customers, (Hermes 2019)).

To the best of our knowledge, we are the first to offer high-quality decision support for companies planning to communicate reliable and narrow time windows to customers while keeping their cost-efficient scheduling and routing algorithms or while outsourcing logistics to third parties. Our work addresses a new variant in the class of problems comprising self-imposed time windows and the time-window-assignment vehicle routing problem, introduced by Jabali et al. (2015) and Spliet and Gabor (2015). We model the problem as a chance-constrained math program. Under mild assumptions about arrival time distributions to customers, we identify structure in an optimal policy that significantly simplifies the optimization. Our main result reduces a high-dimensional stochastic non-linear optimization problem to a root-finding problem in one dimension. For more general distributional forms, the result inspires a practice-ready heuristic. Via a computational study of next-day service operations in a Midwest US city, we show that the heuristic performs competitively relative to a dual bound and significantly outperforms industry standard practices. In short, our work makes complex optimization accessible to large and small firms alike, thus laying the groundwork to improve the customer experience while maintaining a high standard of service.

Through our analytical and computational work, we identify five managerial insights:

1. **Time windows should be customer-specific**, with both size and placement reflecting the shape of the arrival time distribution. Notably, we show that relative to the current practice of communicating uniform time windows to all customers, our method yields a substantial increase in customer convenience without sacrificing reliability of service.
2. We show that **higher service levels require wider time windows** and that **at a given service level, optimization is required to set sizes**.
3. We connect time window size to time of request, showing that in general **firms should assign larger time windows to earlier requests and smaller time windows to later requests**.
4. We tie time window size to location of request, demonstrating that managers should **assign larger time windows to customers further from the depot of operation and smaller time windows to closer customers**.
5. We identify cases where customers are better served by the communication of **two potential service windows** instead of just one.

The paper is organized as follows. In §2, we discuss related literature. In §3, we describe the problem and formulate an optimization model. In §4, we present and analyze our main theoretical result. In §5, we propose a practice-ready heuristic and demonstrate its performance via computational experiments. We conclude the paper in §6.

2. Related Literature

Our research stems largely from literature on the self-imposed time window problem (SITWP), introduced by Jabali et al. (2015), and the time-window-assignment vehicle routing problem (TWAVRP), introduced by Spliet and Gabor (2015). In these streams, the work of Vareias et al. (2019) is most closely related to this paper. Vareias et al. present a variant of the SITWP where all customers are known and the objective is to minimize the sum of time window widths for given routes in the face of stochastic travel times. Each customer's time window is chance-constrained, with probability distributions represented by a small set of discrete scenarios. In contrast, we draw on the industry practice of modeling reliability as a global chance constraint across all customers. Further, our model is general enough to incorporate any source of uncertainty that can be captured in the arrival time distribution, including routing decisions. This allows us to focus on the case where not all requests are known when the time window must be communicated to the customer.

Instead of relying on information about the arrival time distribution, as we do in this paper, Ulmer and Thomas (2018) propose a supervised learning procedure to predict mean arrival times to customers. These predictions are used to set fixed-width time windows for all customers. We benchmark our customer-specific time windows against those of Ulmer and Thomas. On average, our method provides comparable service levels with significantly narrower time windows.

Our analytical results explain observations noted in the literature. When customer choice is a consideration, Köhler et al. (2020) offer some customers a menu of "short" time windows and other customers a selection of "long" time windows, finding it is best to reserve shorter time windows for customers who request later in the operating horizon. Our analytical results explain why the approach of Köhler et al. works, as well as the observation from Ulmer and Thomas (2018) that time window size tends to decrease with time of request: requests that occur later in the booking period correlate with less arrival time variability, hence

these customers can be quoted smaller time windows. However, we note that customers who call later in the booking period in hopes of getting a relatively narrower time window run the risk of receiving no service at all as there may not be capacity to serve them.

The bulk of the SITWP and TWAVRP literature integrates routing and time window assignment in the face of various uncertainties, such as travel time, with the objective of minimizing expected cost. In contrast, our work recognizes that many firms decouple these choices, either because routing efficiency is key or because routing is managed by a third party. This literature includes Madsen et al. (1996), Jabali et al. (2015), Spliet and Desaulniers (2015), Zhang et al. (2015), Dalmeijer and Spliet (2018), Neves-Moreira et al. (2018), Spliet et al. (2018), Subramanyam et al. (2018), Martins et al. (2019), Jalilvand et al. (2021), and Beskers (2022). Additionally, Flinterman (2022) considers a variant of the TWAVRP in which customer locations are known but randomly require service. Hoogeboom et al. (2021) present a robust optimization approach and hedge against time window violations by incorporating arrival time distributions into the optimization via a risk measure. Our work also gives explicit consideration to variability in arrival times. Yu et al. (2023) model a TWAVRP as a two-stage stochastic program. They test their approach in a dynamic environment where customers request sequentially, but these request uncertainties are not considered in the optimization.

Our work contrasts with business models in which customers choose their own service time windows. Waßmuth et al. (2022) provide an overview of this literature, which includes research in grocery delivery (Campbell and Savelsbergh 2005, Ehmke and Campbell 2014, Yang et al. 2014, Köhler et al. 2020, Agatz et al. 2021). Allowing customers to choose their own time windows creates challenges for firms, often preventing the use of efficient scheduling and routing procedures and significantly increasing cost (Ehmke and Campbell 2014). For this reason, particularly in the service contexts that motivate our work, the business model involves a booking period that allows for the planning of efficient routes before the day of service. Related, many grocery companies now set time windows internally instead of via customer preferences (Visser and Savelsbergh 2019).

Finally, our work is related to research on due date assignment, outlined by Keskinocak and Tayur (2004). In this literature, the key decision is to identify a deadline, rather than a time window, by which a product

will be manufactured, with uncertainties in production lead times typically modeled via queuing systems. Similar to our approach, the work of Spearman and Zhang (1999) identifies necessary conditions for optimal deadlines. In the case of setting time windows, we show that necessary conditions are also sufficient. Further, in contrast to Spearman and Zhang, we derive a heuristic from our analytical results and show that it performs well relative to distributions that do not satisfy the assumptions of our analysis. Hafızoğlu et al. (2016), Kahvecioglu and Balcıoglu (2016), and Gel et al. (2020) summarize more recent work on due date management, which considers deadline setting in tandem with pricing decisions and demand conditional on quoted due dates. In contrast, our work focuses on service settings where customers usually are not given a choice regarding the time of service and we consider the more complex task of selecting time windows.

3. Problem Description and Model

In this section, we describe and model the problem of setting service time windows in the face of arrival time uncertainty. We consider the case of a company that provides attended home service to customers. For example, appliance installation and repair, furniture delivery and setup, parcel delivery, utility service, or home health assistance. We assume that the company collects customer requests over a period of time and then provides service on a specified future date. This setting is similar to that of Apte et al. (2007). When a customer requests service, the planner consults the firm's operational state to determine if the request can be accommodated. The state may contain information on customers the firm has already committed to serve, available resource capacities, or a tentative service plan. Before communicating a time window to the customer, the planner considers how arrival time uncertainties may impact service. Between the time of request and the end of the planning horizon, various events may occur, e.g. additional requests, updates to capacity requirements, or handoff to a third party for external logistics planning. In addition, during the execution of service, the technician may encounter other uncertainties such as traffic and unforeseen service challenges. These trajectories of possible events characterize an arrival time distribution, conditional on the operational state. The planner seeks a policy that minimizes expected time window size subject to a constraint on the expected likelihood of early or late arrivals across possible states. For example, the time windows are such that on average 95 percent of customers are served in their given time window (Hermes

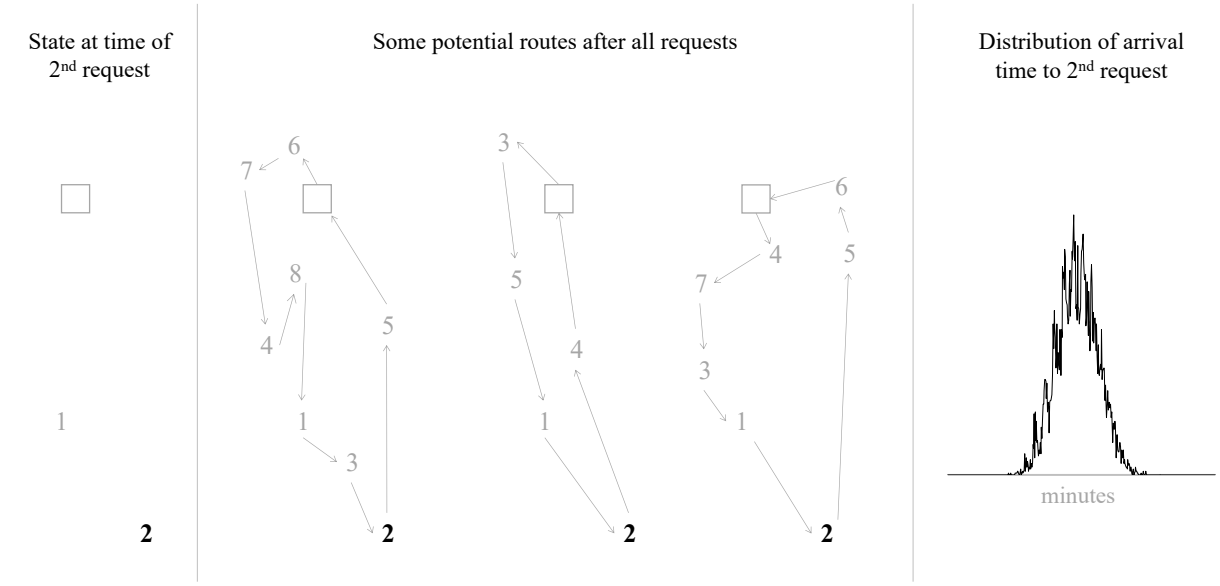

Figure 1 Impact of Logistics Planning on Arrival Times

2019, Weiss 2012). We assume that customers are available for service when the technician arrives, perhaps as a result of a more precise arrival time being communicated on the day of service as was described in the introduction.

Figure 1 illustrates how stochasticity in customer requests and a firm's logistical methods can lead to uncertain arrival times to customers. The left panel shows the operational state at the time of the second request for service on a given day. At this moment in time, the locations of Customers 1 and 2 are known, but the times and locations of any future requests are unknown, as is the sequence of visits in a final routing plan. The middle panel shows several potential route plans after all requests have been realized. The routes vary widely depending on the locations of additional requests. Because Customer 2 is sequenced fourth, fifth, and seventh across the three route plans, the realized arrival time to Customer 2 is different in each plan. Many other realizations of requests and route plans may be possible. The collection of these realizations, along with their respective likelihoods, account for the bulk of the uncertainty in arrival times (Ulmer and Thomas 2018). The result is a distribution of possible arrival times. The right panel displays the results of a simulation to illustrate a potential distribution of arrival times to Customer 2.

Whatever routing and scheduling method the firm uses, or whatever procedure a third-party logistics provider employs, we take it as exogenous to the model, accounting for them only through the arrival time

distributions that they induce. Thus, we assume the final service plan is determined independently from time windows, and thus time window choices do not impact arrival time distributions. This reflects the perspective taken by many firms, that early or late arrivals, no matter their frequency and magnitude, are an acceptable way to reconcile differences in what is initially communicated to customers and an eventual route plan. Because these firms typically view an early or late arrival as a no-cost recourse, time window decisions have no impact on logistical choices. Consequently, whether the routing mechanism is as simple as a cheapest insertion heuristic or as sophisticated as solving a complex mixed integer program, the result is an arrival time distribution associated with the operational state at the time of a customer's request.

Because time window choices do not impact arrival time distributions, the optimization is not a sequence of interdependent decisions. Time window choices are only connected via a global constraint on service level. However, the optimization is still complex. We must choose a time window for every state-dependent arrival time distribution. Minimizing expected time window duration subject to a service level requirement is nonlinear and non-convex in both the objective and the constraint.

Formally, let S be the random state of the environment when the customer requests service, with support $\{s_1, s_2, \ldots, s_n\}$, and denote by $X(s)$ the random arrival time when the state of the environment is s. In this way, we give consideration not only to the distribution of operational states, but also to the arrival time distribution associated with each state. Then, for a given realization s of S, the firm chooses a range $[l(s), u(s)]$ as a state-dependent service time window, where $l(s)$ and $u(s)$ represent times during the day of service. The firm's objective is to minimize expected time window size across all states of the environment, $\mathbb{E}_S[u(S) - l(S)]$, thus seeking time windows most convenient for customers. The firm imposes a single service level across all states. The state-wise contribution to service is the probability of $X(s)$ lying within the communicated time window. Taking expectation across all states, the expected service level is $\mathbb{E}[\mathbb{P}\{l(s) \leq X(s) \leq u(s) | S = s\}]$, which we require to be at least $1 - \alpha$, where the firm chooses α in the range $[0, 1]$. This constraint limits the number of early or late arrivals to be below a threshold, thus protecting the firm's reputation and ensuring a level of reliability. In summary, the problem we seek to solve is

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_S\Big[u(S) - l(S)\Big] \\ \text{subject to} \quad & \mathbb{E}_S\Big[\mathbb{P}\left\{l(s) \le X(s) \le u(s) \middle| S = s\right\}\Big] \ge 1 - \alpha. \end{aligned} \tag{1}$$

4. Optimization and Analysis

In this section, we characterize the optimal policy. Under mild assumptions, our main result identifies structure that reduces problem (1) from a high-dimensional stochastic non-linear optimization problem to a root-finding problem in one dimension. We then discuss managerial implications of the policy structure.

For the purposes of analysis, we assume each $X(s)$ has non-negative support, each cumulative density function (CDF) $F_{X(s)}$ is differentiable, and each probability density function (PDF) $f_{X(s)}$ is unimodal. Regarding the latter, we require each $f_{X(s)}$ to be strictly monotonic either side of its mode $\bar{x}(s) = \arg\max_x f_{X(s)}(x)$. Denote by $\underline{f}_{X(s)}$ the strictly increasing values of $f_{X(s)}$ left of $\bar{x}(s)$ and by $\overline{f}_{X(s)}$ the strictly decreasing values of $f_{X(s)}$ right of $\bar{x}(s)$. Denote the respective antimodes by $\tilde{x}(s) = \arg\min_x \underline{f}_{X(s)}(x)$ and $\hat{x}(s) = \arg\min_x \overline{f}_{X(s)}(x)$. Examples of such distributions include triangular random variables with non-negative support, truncated normal random variables with non-negative support, and gamma random variables with a shape parameter greater than one.

Lemma 1 and Theorem 1 characterize the optimal policy. When $\alpha < 1$, Theorem 1 shows that each time window shares a common density value $y^\star$ among all PDFs, and that $y^\star$ uniquely sets the service level equal to $1 - \alpha$. Thus, for each state s, optimal time windows are obtained via an inverse mapping of $y^\star$ through $f_{X(s)}$. When $\alpha = 1$, then any solution where all time windows are of zero-width is optimal. This structure follows from an analysis of necessary conditions for optimality, which we show are also sufficient. The role of Lemma 1 is to address boundary conditions required for the proof of Theorem 1.

Lemma 1 (Zero- and Maximum-Width Time Windows) *Choose $y \ge 0$ such that $\mathbb{E}_S[\mathbb{P}\{l(s,y) \le X(s) \le u(s,y) \big| S = s\}] = 1 - \alpha$. For each realization s of S such that $\underline{f}_{X(s)}(\tilde{x}(s)) < y < f_{X(s)}(\bar{x}(s))$, require $l(s,y) = \underline{f}^{-1}_{X(s)}(y)$. If $\overline{f}_{X(s)}(\hat{x}(s)) < y < f_{X(s)}(\bar{x}(s))$, require $u(s,y) = \overline{f}^{-1}_{X(s)}(y)$. Then, the following assignments minimize expected time window size:*

(i) $l(s,y) = \tilde{x}(s)$ *if* $y \leq \underline{f}_{X(s)}(\tilde{x}(s))$,

(ii) $u(s,y) = \hat{x}(s)$ *if* $y \leq \overline{f}_{X(s)}(\hat{x}(s))$,

(iii) and $u(s,y) - l(s,y) = 0$ *if* $y \geq f_{X(s)}(\bar{x}(s))$.

Proof. The proof is by contradiction. Suppose condition (i) does not minimize expected time window size. Then, for any state s in S, when $y \leq \underline{f}_{X(s)}(\tilde{x}(s))$, choosing $l(s,y) < \tilde{x}(s)$ or $l(s,y) > \tilde{x}(s)$ must result in a feasible assignment with a smaller objective value. Yet, in the first case the objective value increases and in the second case the service level constraint is not satisfied, both of which contradict the assumed optimality of the new assignment. A similar argument proves condition (ii). Suppose condition (iii) does not yield a time window assignment with minimum expected size. Then, for any state s in S, when $y \geq f_{X(s)}(\bar{x}(s))$, there must exist $u(s,y)$ and $l(s,y)$ such that the difference is positive, the assignment is feasible, and the objective value is smaller. However, by assumption, a zero-width time window is feasible. Because a zero-width assignment also decreases the objective value relative to a positive-width assignment, the assumed optimality of the positive-width assignment is contradicted. □

Theorem 1 (Optimal Time Windows) *For any* $y \geq 0$*, and for each realization* s *of* S*, let*

$$l(s,y) = \begin{cases} \tilde{x}(s), & y \leq \underline{f}_{X(s)}(\tilde{x}(s)), \\ \underline{f}^{-1}_{X(s)}(y), & \underline{f}_{X(s)}(\tilde{x}(s)) < y < f_{X(s)}(\bar{x}(s)), \\ \bar{x}(s), & \textit{otherwise}, \end{cases} \tag{2}$$

$$u(s,y) = \begin{cases} \hat{x}(s), & y \leq \overline{f}_{X(s)}(\hat{x}(s)) \\ \overline{f}^{-1}_{X(s)}(y), & \overline{f}_{X(s)}(\hat{x}(s)) < y < f_{X(s)}(\bar{x}(s)), \\ \bar{x}(s), & \textit{otherwise}. \end{cases} \tag{3}$$

Then, when $\alpha < 1$*, an optimal solution corresponds to the unique* $y^\star \geq 0$ *satisfying*

$$\mathbb{E}_S\left[\mathbb{P}\left\{l(s,y^\star) \leq X(s) \leq u(s,y^\star) \middle| S = s\right\}\right] = 1 - \alpha. \tag{4}$$

When $\alpha = 1$*, any solution such that* $l(s) = u(s)$*, for each realization* s *of* S*, is optimal.*

Proof. The result follows from the Karush-Kuhn-Tucker (KKT) necessary conditions for constrained optimization. Let $l = (l(s_i))_{i=1}^n$ and $u = (u(s_i))_{i=1}^n$. Then, writing out expectations and using arrival time CDFs, the Lagrangian is

$$L(l,u,\lambda) = \sum_{i=1}^{n} \mathbb{P}\{S = s_i\}\,[u(s_i) - l(s_i)] + \lambda\left[1 - \alpha - \left(\sum_{i=1}^{n} \mathbb{P}\{S = s_i\}\left[F_{X(s_i)}(u(s_i)) - F_{X(s_i)}(l(s_i))\right]\right)\right], \quad (5)$$

where $\lambda \geq 0$ is a scalar. Dual feasibility requires the partial derivatives of the Lagrangian with respect to each $l(s_i)$ and $u(s_i)$ equal zero. For $i = 1, 2, \ldots, n$, we have

$$\frac{\partial L}{\partial l(s_i)} = -\mathbb{P}\{S = s_i\} + \lambda \mathbb{P}\{S = s_i\} f_{X(s_i)}(l(s_i)) = 0 \quad (6)$$

$$\implies f_{X(s_i)}(l(s_i)) = 1/\lambda \quad (7)$$

$$\implies l(s_i) = f^{-1}_{X(s_i)}(1/\lambda) \quad (8)$$

and

$$\frac{\partial L}{\partial u(s_i)} = \mathbb{P}\{S = s_i\} - \lambda \mathbb{P}\{S = s_i\} f_{X(s_i)}(u(s_i)) = 0 \quad (9)$$

$$\implies f_{X(s_i)}(u(s_i)) = 1/\lambda \quad (10)$$

$$\implies u(s_i) = f^{-1}_{X(s_i)}(1/\lambda). \quad (11)$$

For a given state s_i, these equations require $l(s_i)$ and $u(s_i)$ achieve the same density value $1/\lambda$. Thus, either $l(s_i) = u(s_i)$ or $l(s_i) = \underline{f}^{-1}_{X(s_i)}(1/\lambda)$ and $u(s_i) = \overline{f}^{-1}_{X(s_i)}(1/\lambda)$. The second cases in Equations (2) and (3) follow from these requirements, where $y = 1/\lambda$.

Complementary slackness requires the product of λ and the slack in the service level constraint be zero: $\lambda[1 - \alpha - (\sum_{i=1}^n \mathbb{P}\{S = s_i\}[F_{X(s_i)}(u(s_i)) - F_{X(s_i)}(l(s_i))])] = 0$. It follows that $\lambda = 0$, the service level

constraint is binding, or both. By the dual feasibility conditions given in the second cases of Equations (2) and (3), as λ approaches zero, all time window widths go to zero, which can only be feasible if $\alpha = 1$, in which case the slack in the service level constraint is zero. Thus, regardless of the value of λ, the service level constraint must be binding. Consequently, for any realization s of S, when $1/\lambda$ lies outside the codomain of $\underline{f}_{X(s)}$ or $\overline{f}_{X(s)}$, Lemma 1 establishes the first and third cases in Equations (2) and (3), where again $y = 1/\lambda$.

Combining the above with the requirement of primal feasibility, the KKT conditions can be expressed in a single equation:

$$\sum_{i=1}^{n} \mathbb{P}\{S = s_i\} \left[F_{X(s_i)}(u(s_i, y)) - F_{X(s_i)}(l(s_i, y))\right] = 1 - \alpha. \tag{12}$$

Any $y \geq 0$ satisfying Equation (12), and equivalently any $y \geq 0$ satisfying Equation (4), achieves the necessary conditions for optimality.

Finally, examine the shape of the left-hand side of Equation (12). Consider the contribution of a given state s_i. Let $\check{y}(s_i) = \min f_{X(s_i)}$ be the smallest density value across the corresponding PDF. Without loss of generality, assume $\check{y}(s_i) = f_{X(s_i)}(\tilde{x}(s_i)) \leq f_{X(s_i)}(\hat{x}(s_i))$. When $y \leq \check{y}(s_i)$, the time window $(l(s_i, y), u(s_i, y))$ spans the support of $X(s_i)$ and the contribution to the service level is as large as possible. Then, as y increases beyond $\check{y}(s_i)$, the contribution strictly decreases up to a point, then goes to zero. To see this, note that because $\underline{f}_{X(s_i)}$ strictly increases, $\underline{f}^{-1}_{X(s_i)}$ strictly increases, and because $\overline{f}_{X(s_i)}$ strictly decreases, $\overline{f}^{-1}_{X(s_i)}$ strictly decreases. Thus, when $\check{y}(s_i) \leq y \leq f_{X(s_i)}(\hat{x}(s_i))$, $F_{X(s_i)}(l(s_i, y))$ strictly increases with y and $F_{X(s_i)}(u(s_i, y))$ is constant. Then, when $f_{X(s_i)}(\hat{x}(s_i)) \leq y \leq f_{X(s_i)}(\bar{x}(s_i))$, $F_{X(s_i)}(l(s_i, y))$ strictly increases with y and $F_{X(s_i)}(u(s_i, y))$ strictly decreases with y. In both cases, the difference $F_{X(s_i)}(u(s_i, y)) - F_{X(s_i)}(l(s_i, y))$ strictly decreases with y. Then, when $y > f_{X(s_i)}(\bar{x}(s_i))$, the difference is constant at zero. If $\check{y}(s_i) = f_{X(s_i)}(\hat{x}(s_i))$, the contribution of state s_i can be similarly characterized. It follows that the left-hand side of Equation (12) is a weighted sum of functions, each of which begins at some value, then strictly decreases up to some y, and then goes to zero.

Let $\check{y} = \min_{i=1}^{n} \check{y}(s_i)$ be the smallest density value across all PDFs and denote by $\bar{y} = \max_{i=1}^{n} f_{X(s_i)}(\bar{x}(s_i))$ the largest density value across all modes. Then, the service level strictly decreases as y increases from $\check{y}$ to $\bar{y}$,

then goes to zero. Thus, when $\alpha < 1$, y must be less than $\bar{y}$, and there is a unique y satisfying Equation (12). In this case, because only one solution satisfies the necessary conditions for optimality, the conditions are also sufficient. When $\alpha = 1$, any solution with all time window widths set to zero satisfies Equation (12). In this case, because all such solutions have equivalent objective values of zero, each is optimal. □

We graphically describe Theorem 1 via Figure 2, where we illustrate the optimization across only two equally-likely states s_1 and s_2. The left portion of the figure illustrates how time windows change with the common density value. The figure shows the arrival time distributions for random variables $X(s_1)$ and $X(s_2)$, which we represent with triangular distributions for ease of illustration. PDF $f_{X(s_1)}$ has a lower limit of 5, an upper limit of 11, and is symmetric about its mode $\bar{x}(s_1) = 8$. PDF $f_{X(s_2)}$ is positively skewed about its mode $\bar{x}(s_2) = 17$ with a lower limit of 16 and an upper limit of 20. The mean, standard deviation, and skewness values are displayed at bottom. Three density values are depicted to facilitate subsequent discussion.

At density value $y_1 = 0.09$ (which corresponds to an industry-standard service level of 0.95 (Hermes 2019)), the time windows $[l(s_1, y_1), u(s_1, y_1)] = [5.8, 10.2]$ and $[l(s_2, y_1), u(s_2, y_1)] = [16.2, 19.5]$ are wide, near the upper and lower boundaries of the supports. Increasing the density level to $y_2 = 0.20$ (which corresponds to a service level of 0.74), time windows $[l(s_1, y_2), u(s_1, y_2)] = [6.8, 9.2]$ and $[l(s_2, y_2), u(s_2, y_2)] = [16.4, 18.8]$ have equal width and are smaller relative to their sizes at y_1. Because density level $y_3 = 0.42$ (corresponding to an impractically low service level of 0.15) is larger than $f_{X(s_1)}(8) = 0.33$, $l(s_1, y_3) = u(s_1, y_3) = 8$ and the time window size is zero. Time window $[l(s_2, y_3), u(s_2, y_3)] = [16.8, 17.5]$ is even narrower than it is at y_2. Although service levels as low as those corresponding to y_2 and y_3 are uncommon in practice, they are helpful in the illustration of Theorem 1.

The top-right portion of Figure 2 graphs Equation (4), depicting the relationship between the common density value and the service level. As the density value increases from y_1 to y_2 to y_3, the service level decreases from 0.95 to 0.74 to 0.15, respectively. The kinks in the graph occur at $f_{X(s_1)}(8) = 0.33$ and $f_{X(s_2)}(17) = 0.5$, the density values of the modes, which push their respective time windows to a size of zero. As illustrated in the proof of Theorem 1 and reflected in the figure, the left-hand side of Equation (4) is strictly

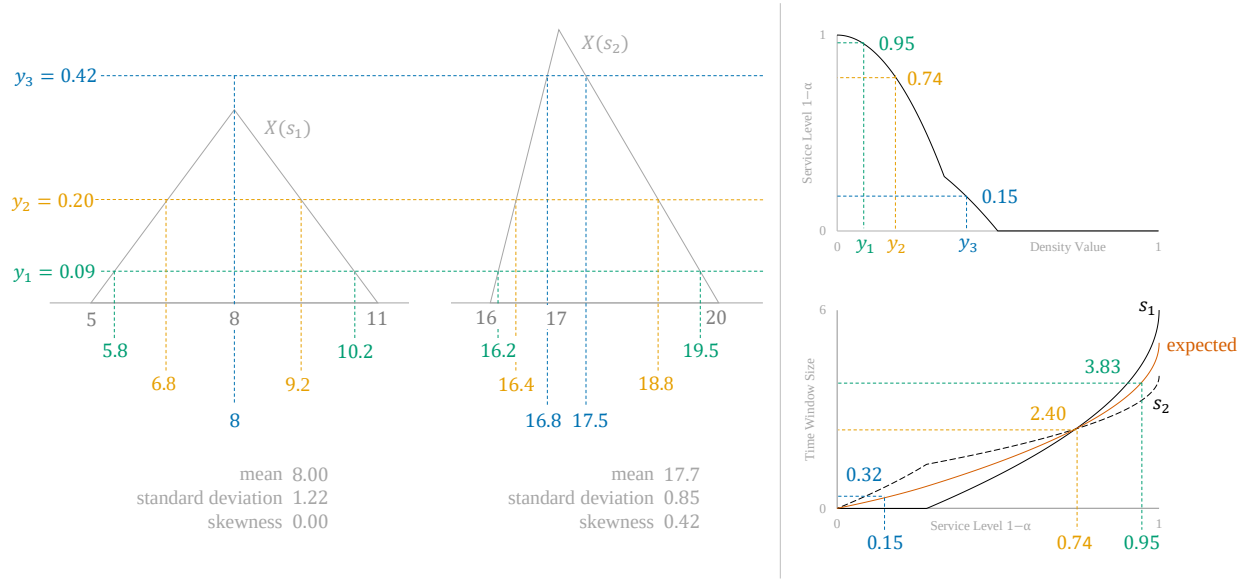

Figure 2 Time Windows, Service Levels, and Common Density Values

Note. As the common density value increases, time windows shrink (left), the service level goes down (top-right), and the minimum expected time window size associated with each decreases as do individual time windows (bottom-right).

decreasing. Thus, basic root-finding methods, such as bisection, can be used to identify the point at which the function equals $1-\alpha$, thereby identifying $y^\star$ and allowing time windows to be set per Equations (2) and (3).

Via service level, the bottom-right portion of Figure 2 illustrates the connection between common density values and objective values. As the service level increases from 0.15 to 0.74 to 0.95, the minimum expected time window sizes move from 0.32 to 2.40 to 3.83, respectively, corresponding to smaller common density values y_3, y_2, and y_1. The figure also shows how the time window size for each state contributes to the objective across the same domain of service levels. Thus, for a given service level, an optimal solution to problem (1) can be identified by finding the common density value satisfying Equation (4).

Two managerial insights follow directly from Theorem 1. First, in contrast to both practice and the academic literature, time window size and placement should reflect the shape of the arrival time distribution. The left portion of Figure 2 illustrates this, where the skewness of the arrival time distributions, in conjunction with the common density value, dictate the size and placement of time windows. The popular industry practice of fixed-width time windows suggests customer arrival time distributions are ignored in favor of assigning customers to four-hour morning or afternoon buckets. This practice should be retired in favor of customer-specific time windows tailored to arrival time distributions.

Ulmer and Thomas (2018) consider the time-dependent time windows, but center them around expected arrival times. According to Theorem 1, this can only be optimal when the arrival time distribution is symmetric about its mean, as illustrated by the arrival time distribution for state s_1 in Figure 2. However, in state s_2, and in any state whose arrival time distribution is not symmetric about its mean, the time window is not centered about the expected arrival time. Indeed, in the extreme case of a 0.15 service level (corresponding to common density value y_3), not only is the optimal time window $[l(s_2, y_3), u(s_2, y_3)] = [16.8, 17.5]$ not centered about the mean of 17.7, the time window does not even include the mean. As our numerical experiments in the subsequent section show, customer-specific time windows that follow the shape of the arrival time distribution increase convenience without sacrificing reliability. We summarize these observations in Insight 1.

Insight 1 (Customer-Specific Time Windows) *The beginning and end of a time window should reflect the shape of the arrival time distribution. Consequently, the practices of assigning uniform fixed-width time windows to all customers and of centering time windows around mean arrival times should be replaced by assignment of customer-specific time windows.*

Second, intuition suggests, and Theorem 1 confirms, that higher service levels require wider time windows to better hedge against uncertain arrival times. However, at a given service level, the assignment of time windows is not necessarily dictated by the uncertainties, as one might assume. For example, focusing on the bottom-right of Figure 2, when the service level is 0.74, the optimal policy assigns equally-sized time windows to each state s_1 and s_2: $9.2 - 6.8 = 18.8 - 16.4 = 2.40$. At higher service levels, where most businesses operate, state s_1, whose arrival time standard deviation is higher than that of state s_2, receives the larger time window. This satisfies the intuition that the arrival time distribution with more uncertainty warrants the wider service window, an inclination largely confirmed by our numerical experiments in the subsequent section. However, at lower service levels, the larger time window goes to state s_2. In fact, at service levels less than or equal to 0.28, state s_1 is assigned a zero-width time window, a commitment that almost certainly cannot be kept. Though this scenario is a concern in principle, our experiments in §5 show that such extreme cases are rare when operating at industry-standard service levels and that early or late arrivals are infrequent and small in magnitude.

This example highlights the importance of optimization. Not only does this behavior defy managerial instinct, it is not reflected in industry practice. As discussed in §1, the home services industry assigns fixed-width time windows to customers, often four hours in duration. Further, much of the TWAVRP literature reviewed in §2 assumes fixed-width time windows. Though Köhler et al. (2020) and Ulmer and Thomas (2018) provide computational evidence that differentiated time windows are beneficial, Theorem 1 shows this practice to be optimal and provides guidance on setting time window sizes proportional with the service level. We formalize these observations as Insight 2 and further probe the relationship between arrival time uncertainty and time window size in §5.

Insight 2 (Time Window Size and Service Level) *Higher service levels require wider time windows. At a given service level, optimization is required to size time windows.*

5. Theory to Practice

Though the assumptions leading to Theorem 1 allow for an elegant analysis, they are not always satisfied in practice. For example, in our computational study, we encounter a variety of empirical PDFs, from approximately unimodal to bimodal with high variance. Examples are given in Figure 3, where the horizontal axes represent arrival time realizations for three states across a 480-minute operating horizon and the vertical axes display relative frequencies. The vertical axes vary in scale to aid illustration. The common density value $y^\star = 0.0037$ displayed in the figure aids subsequent discussion. Though many of the empirical PDFs we observe resemble the left-most portion of Figure 3, even in these cases, values do not always strictly decrease left and right of the mode, as assumed in the analysis of §4. Thus, for a given density value y, how to apply Theorem 1 via Equations (2) and (3) is unclear because more than two elements of the support achieve the value.

In §5.1, we present a heuristic algorithm, motivated by the structure of Theorem 1, that is applicable to general arrival time distributions. §5.2 describes the simulation designed to demonstrate the quality of the heuristic. We specify benchmarks in §5.3 followed by a discussion in §5.4.

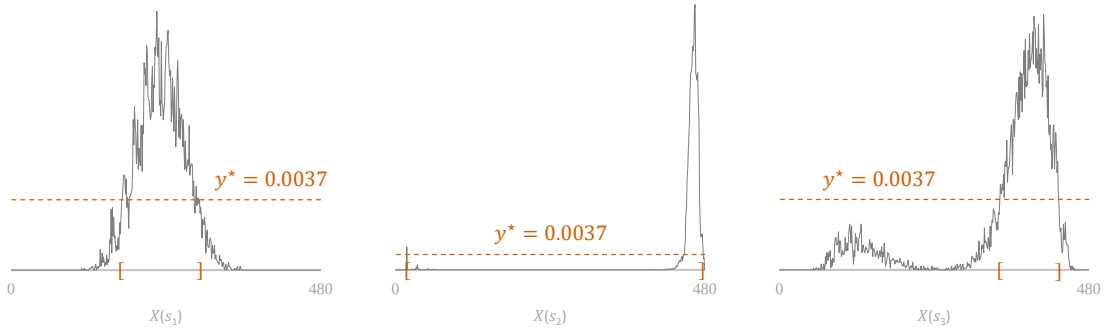

Figure 3 Empirical PDFs and Time Windows

Note. Left: approximately unimodal. Center: low variance, large time window. Right: high variance, small time window. Vertical scales vary to aid illustration.

5.1. Heuristic Solution Approach

For state s and some common density value y, denote by $\mathcal{X}(s,y) = \{x : f_{X(s)}(x) = y\}$ the set of support elements whose density values are y. In the case of discrete distributions (e.g., minute-by-minute), assume support elements are indexed in ascending order and let $\mathcal{X}(s,y) = \{x_i(s) : f_{X(s)}(x_{i-1}(s)) \leq y \leq f_{X(s)}(x_{i+1}(s))\}$ be the set of support elements $x_i(s)$ such that y lies between the masses associated with $x_{i-1}(s)$ and $x_{i+1}(s)$. Then, select the corresponding time windows by setting $l(s,y)$ to the minimum value in $\mathcal{X}(s,y)$ and $u(s,y)$ to the maximum value. The use of the common density value is analogous to the optimal solution given in Theorem 12 that identifies a unique $y^\star$ as optimal across all distributions. Given this mapping of a common density value to time windows, we proceed as in §4 and search for the smallest y such that the service level constraint is satisfied.

Algorithm 1 formalizes the heuristic. The procedure takes as input a step size η on Line 1, a value d representing the maximum density value across all distributions, and initializes the common density value to $y = d$ on Line 2. In our computational study, we set $\eta = 0.0001$, which corresponds to the granularity of the empirical PDFs constructed via simulation. Parameter η can also be viewed as a discretization for continuous-time distributions. Each iteration of the loop beginning on Line 3 sets time windows and checks feasibility. For each state s in $\{s_1, s_2, \ldots, s_n\}$, Lines 4-8 map y to a time window $[l(s), u(s)]$. If y lies below the density value of the mode, $f_{X(s)}(\bar{x}(s))$, then Line 6 makes the assignment as described above, else Line 8

sets $l(s)$ and $u(s)$ to the mode. Line 9 increments y by η and Line 10 checks the service level of the current time windows against the threshold $1-\alpha$. Once the service level is high enough, the procedure terminates on Line 11 with time windows for each state.

Algorithm 1 Time Window Heuristic

1: **input:** step size η
2: $y \leftarrow d$
3: **repeat**
4: **for all** $s \in \{s_1, s_2, \ldots, s_n\}$ **do**
5: **if** $y < f_{X(s)}(\bar{x}(s))$ **then**
6: $l(s) \leftarrow \min \mathcal{X}(s,y)$ and $u(s) \leftarrow \max \mathcal{X}(s,y)$
7: **else**
8: $l(s) \leftarrow \bar{x}(s)$ and $u(s) \leftarrow \bar{x}(s)$
9: $y \leftarrow y - \eta$
10: **until** $\mathbb{E}_S\left[\mathbb{P}\left\{l(s) \leq X(s) \geq u(s) \middle| S = s\right\}\right] \leq 1-\alpha$
11: **output**: time windows $([l(s_i), u(s_i)])_{i=1}^n$

Figure 4 depicts the results of the heuristic on the data discussed in §5.2. Even when the assumptions surrounding Theorem 1 are not met, we observe a strict decrease in service level as a function of the common density value, as shown in the left portion of the figure. When the service level is set to $1-\alpha = 0.95$, Algorithm 1 terminates with common density value $y^\star = 0.0037$, achieving an objective value of 77.3 minutes, as shown in the right portion of the figure. Thus, though Algorithm 1 conducts a simple search, Figure 4 suggests this is sufficient. Further, as we show in §5.4, the time windows selected by Algorithm 1 perform very well.

5.2. Simulation Study

For our computational study, we use the same simulation environment introduced in Ulmer and Thomas (2018). We refer the reader to that paper for details. The setting for our computational study is service routing across the Iowa City area, a region in the Midwest United States with a population of approximately 160,000 and more than 32,000 residential locations, with a hub of operation situated on the outskirts. We consider

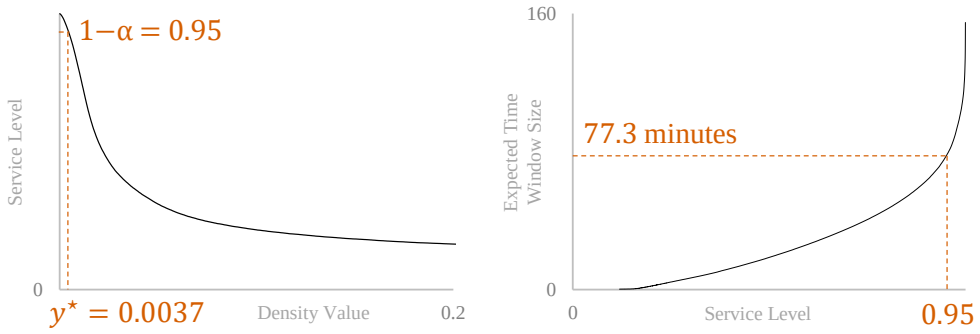

Figure 4 Algorithm 1 Mimics Analytical Results

the case of next-day service with a two-day business model. Requests are collected during an eight-hour capture phase on day one and services are carried out across an eight-hour execution phase on day two with service times of 15 minutes at each customer. Service time windows are communicated at the time of request, and routes are constructed and adjusted on-the-fly throughout the capture phase. Thus, in the context of our optimization model, a realization s of random state S consists of the time of request, the requesting customer's location, and the set of customers the firm has already committed to serve, all of which instantiate an arrival time distribution $X(s)$ for delivery during the day-two execution phase.

We employ sampling to generate state and arrival time realizations. Specifically, across one-minute time increments, we simulate a stochastic request process alongside an insertion-based routing mechanism. A request is accepted if it can be served within the operating horizon. Otherwise, it is rejected for next-day service, which in practice typically means postponement of service to the following day. An accepted request triggers generation of a state. We generate $\tilde{n} = 10,000$ states $\{s_1, s_2, \ldots, s_{\tilde{n}}\}$ representing approximately 500 days of operation. Then, for each day-one sampled state s, we estimate the day-two arrival time distribution via an additional $\hat{n} = 10,000$ simulations. Each trajectory begins from state s during day one, generates new requests through the end of the capture phase, and routes these requests. The resulting delivery sequence determines the next-day arrival time for the customer in state s. Denote the simulated arrival times as $\{\hat{x}_1(s), \hat{x}_2(s), \ldots, \hat{x}_{\hat{n}}(s)\}$, which we use to construct an empirical PDF for $X(s)$, e.g., as in the examples of Figure 3.

In addition to building empirical PDFs via simulation, we utilize the simulations as input to estimate policy values and service levels. Given time windows $([l(s_i), u(s_i)])_{i=1}^{\tilde{n}}$, we estimate the objective value as

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} u(s_i) - l(s_i). \tag{13}$$

We estimate the corresponding service level as

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} 1\Big\{\hat{x}_j(s_i) \in [l(s_i), u(s_i)]\Big\}, \tag{14}$$

where $1\{\cdot\}$ is an indicator function returning one if the arrival time falls within the service time window and zero otherwise. In our study, we set $\alpha = 0.05$ to correspond with common industry practice of 95 percent on-time service (Hermes 2019).

5.3. Benchmarks

We gauge our method for setting service time windows against a dual bound and three policies. To assess the effectiveness of the heuristic described in Algorithm 1, we calculate a Lagrangian dual bound. Specifically, we employ subgradient optimization to maximize the minimum value of the Lagrangian in Equation (5), where expected time window size and service level are estimated as described in §5.2. In the absence of a guarantee of an optimal policy, this approximate dual bound serves as an absolute benchmark, small gaps indicating the method is likely good enough for practice.

Our second benchmark explores the effectiveness of setting time windows via Algorithm 1 when information is limited. For example, when it is not possible to explicitly estimate the arrival time distribution, one may have to rely on moment information to perform the estimation. For each state, we use the first two moments from our arrival time simulations to estimate parameters for a symmetric triangular arrival time distribution, which we use to set time windows via Equations (2) and (3), estimating objective values and service levels via simulation, as described in §5.2. We refer to this policy as "Limited" and note that experiments with truncated normal distributions yield comparable performance.

The third benchmark exploits an observation from the literature that time window size tends to decrease with time of request (Ulmer and Thomas 2018, Köhler et al. 2020). Motivated by this, we center a time window around the mean of the empirical PDF, then set the width per a decreasing linear function. We require time windows be as small as zero minutes, no larger than 480 minutes, and lie within the 480-minute

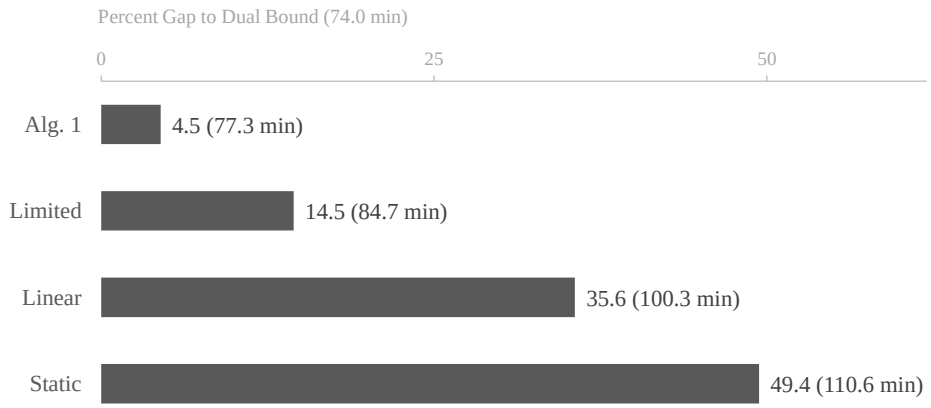

Figure 5 Comparison of Policy Values

operating horizon. Thus, some time windows may be truncated. We adjust the slope to achieve the smallest expected time window size such that the service level is at least $1-\alpha$. As before, we estimate objective and constraint values via simulation. We refer to this policy as "Linear."

Finally, the fourth benchmark mirrors current practice by offering a fixed-size service time window to each customer (Weiss 2012). Centering time windows around means, and requiring time windows to begin and end within the 480-minute operating horizon, we identify the smallest range such that the service level is at least $1-\alpha$, simulating to estimate objective and constraint values. We denote this policy as "Static."

5.4. Discussion

Our discussion examines policy performance across various metrics of managerial interest related to convenience and reliability: time window size, time window violation, and how arrival time variability, time of request, and location of request impact time windows. Figure 5 displays policy quality as measured by the objective and relative to the Lagrangian dual bound. For each policy, the figure shows the gap as a percentage of the dual bound. Notably, setting time windows via Algorithm 1, as depicted in Figure 4, results in a gap of less than five percent, or just over three minutes. Even with limited information, the expected time window size increases by less than eight additional minutes. Although the requirements for Theorem 1 to yield an optimal policy are not entirely aligned with real-life circumstances, this result demonstrates that the theory transfers well to practice.

In contrast, Figure 5 suggests that choosing time windows as a linear function of time of request performs poorly, with a gap of more than 35 percent relative to the dual bound. Further, our methods generate substantial improvement relative to industry practice. Offering all customers a fixed-width time window is far from optimal, with nearly a 50 percent duality gap and an increase of more than 30 minutes in expected time window size relative to our best method. As all of the policies satisfy the $1 - \alpha$ service-level threshold, these results suggest that firms setting time windows via time of request, as well as firms using static time windows, can significantly increase customer convenience without sacrificing reliability. In particular, the results underscore the importance of Insight 1, that time windows should be customer-specific and tailored to arrival time distributions.

In addition to expected time window size, service quality is an important managerial concern. Much like the classical inventory service level metric captures stockout occurrence without regard to stockout quantity, the service level constraint in Equation (1) explicitly measures whether or not a time window violation occurs, but not the magnitude of the violation. Because the severity of an early or late arrival has managerial importance, Figure 6 characterizes for each policy the distribution of expected time window violations. Each horizontal histogram displays the number of states, across all simulations, for which the expected violation was five minutes or larger. The figure also displays the average and maximum expected violations. For reasons of scale, we do not display frequencies for expected violations of five minutes and less. For each policy, these cases account for at least 93 percent of the simulations.

Figure 6 reveals that the average expected violation is similar for each policy, between 72 and 90 seconds. The policies are more readily differentiated when we look to the maximum expected violations. At worst, Algorithm 1 is 48 minutes outside the communicated time window. In contrast, worst-case values for the other policies range from nearly two hours to nearly three hours. One explanation for the difference in performance is that Algorithm 1 accounts for the skewness of the arrival time distribution, whereas the other policies center time windows about the mean. Thus, while intuition might suggest that the magnitude of violations should increase with decreases in expected time window size, Figure 6 indicates this is not the case. Thus, Algorithm 1 not only promises smaller time windows at a given service level, but posts favorable performance when a commitment is not kept.

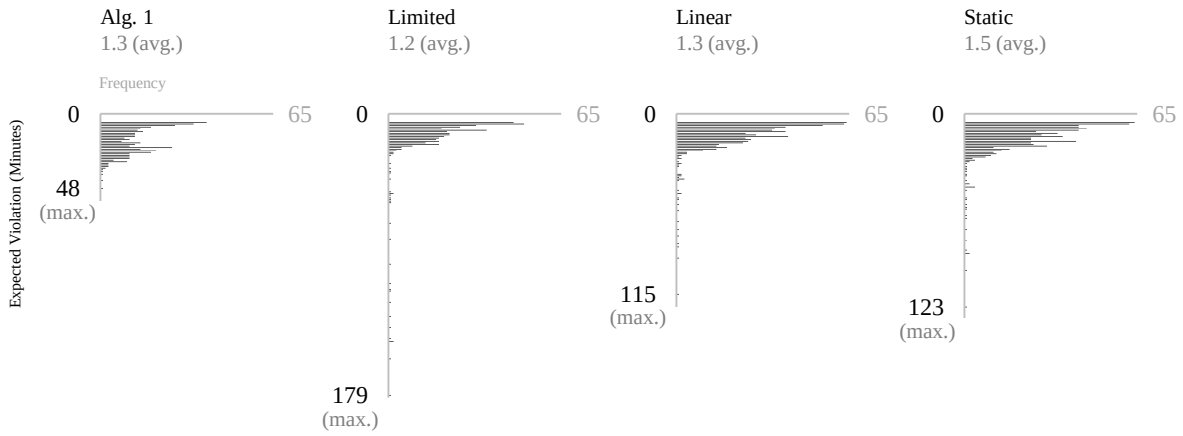

Figure 6 Magnitude of Time Window Violations

Beyond policy performance with respect to convenience and reliability, we seek to better understand the characteristics of the time windows set by Algorithm 1. Figure 7 examines the relationship between variability in arrival time and time window width. The heat map displays results across all simulations, marking areas of high concentration with darker colors and areas of lower concentration with lighter colors. The overall trend suggests that as standard deviation of arrival time increases, time windows widen, generally aligning with the intuition that more uncertainty warrants a larger range for service.

Although Figure 7 shows strong positive correlation between arrival time variability and time window size, a closer look at outliers is instructive. In the bottom-right portion of the figure, we find high-variance distributions assigned small time windows, while in the top-left we find low-variance distributions with large time windows. Circumstances leading to these cases are displayed in Figure 3, the center and right PDFs, respectively. Though the bulk of the probability mass in the center distribution is concentrated at the end of the service horizon, a small portion of the mass is situated toward the beginning. This reflects the strong likelihood of receiving service toward the end of a route, with a small chance of being among the first in the delivery sequence. Despite the center distribution's heavy right skewness, the common density value of $y^\star = 0.0037$ is small enough to capture the left-most mass, thus assigning a nearly eight-hour time window to a low-variance customer. In contrast, though the right-most PDF in Figure 3 is also right-skewed, it is appreciably more bimodal than the center PDF. However, its probability mass is spread such that the chances

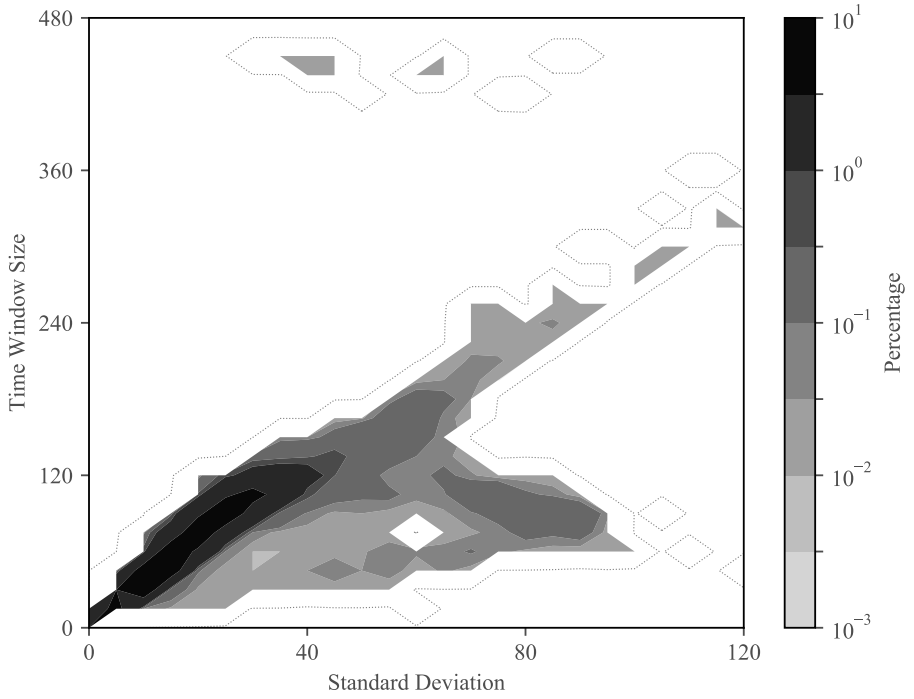

Figure 7 Impact of Arrival Time Variability on Time Window Size

of early arrival are all less than $y^\star = 0.0037$, thus leading to a much smaller time window for a high-variance customer.

Each PDF presents difficulties: either the time window is unnecessarily large or the likelihood of meeting the time window is relatively small, both leading to potential customer dissatisfaction. Following this line of thinking to an extreme case, where arrival time variability is so high that the distribution approaches a uniform PDF, a large enough $y^\star$ will assign a time window width of zero to this ultra-high-variance customer, a commitment that cannot be met. Though this scenario is a concern in principle, our results suggest it occurs primarily when service levels are impractically low, in which case $y^\star$ may even dominate distributions with moderate variability. In the context of lead time quotation, Spearman and Zhang (1999) raise a similar concern, suggesting that different service metrics may be warranted. However, as Figure 7 illustrates, when the $1-\alpha$ threshold is set high per industry standards, the vast majority of customers receive time windows proportional to their arrival time variability. Further, we know from Figure 6 that large time window violations are rare. Thus, these extreme cases are unlikely to lead to poor management practices.

In any case, when time window size and arrival time uncertainty are mismatched, managerial intervention may be necessary to supplement the optimization. To solve the conundrum posed by bimodal arrival time distributions, we suggest communicating two time windows to the customer, one toward the beginning of the operating horizon and a second toward the end. From the customer's perspective, service during either of two smaller time windows is preferable to service during one much larger time window. Similarly, two potential service periods are preferred over initial communication of a small time window followed by a drastic correction. In both scenarios, two time windows reduce the chances of last-minute schedule disruptions due to corrections and affords the customer better opportunity to plan. We summarize this suggestion in Insight 3.

Insight 3 (Time Window Pairs) *Communicate two service time windows to customers whose arrival time distributions are bimodal.*

It is instructive to consider the impact of variability on time window size in conjunction with time of request and the likelihood of the customer's request to be accepted for day-two service. The left side of Figure 8 displays the average time window size across the first six hours of the capture phase (the point at which routes are typically full) when setting time windows via Algorithm 1. The right side shows the percent of requests accepted by time of request across the same period. Taken together, Figure 7 and the left portion of Figure 8 suggest variability in arrival time is negatively correlated with time of request. In other words, customers who request service later in the capture phase often receive a smaller time window than customers who request earlier, largely because routing uncertainty is smaller when the potential for additional customers and further route adjustments is lower. As we will show, request location together with temporal considerations gives a more complete picture of factors influencing time window size.

Although on the surface it may appear there is convenience to be had by requesting later, customers who follow such a strategy may not receive next-day service. The right side of Figure 8 shows that over the first two hours of capture, most requests are accepted. Then, as routes fill and capacity is consumed, later-in-the-day requests cannot always be feasibly serviced, leading to fewer and fewer acceptances as the capture phase progresses. In practice, requests rejected for next-day service would likely be considered early requests for service two days out, and consequently receive a large time window. Thus, though earlier service

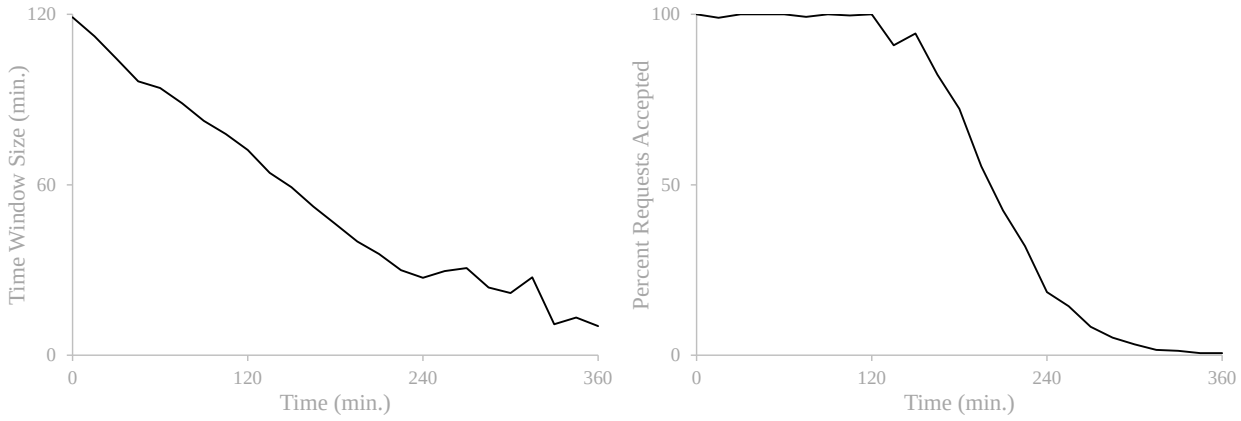

Figure 8 Time Window Sizes and Accepted Requests Across Capture Phase

requests tend to receive larger time windows, likelihood of next-day service can be significantly higher relative to customers who request later.

Initially, differences in time window size by time of request may be cause for concern because customers are not treated equitably with respect to convenience. However, we see similar situations in other aspects of business. Just as late arrival to a store could result in marked down leftovers or a stockout, a later request could lead to a smaller time window, postponement of service, or no inventory left at all. Further, compared to the industry standard static policy, which according to Figure 5 gives each customer a time window of just over 110 minutes, the left side of Figure 8 shows that when time windows are chosen via Algorithm 1, the vast majority of customers receive smaller time windows. Consequently, even when time windows are tailored to individual customers, collective convenience is not only improved, but individual convenience is very often higher as well, all without compromising service. We summarize the managerial implications for setting time window size based on time of request in Insight 4.

Insight 4 (Time of Request) *Assign larger time windows to earlier requests and smaller time windows to later requests.*

Next, we examine the role of location on time window size. Figure 9 displays the Iowa City area on an 10x10 grid with the delivery depot represented as a red circle in the upper-left. The heat map displays time

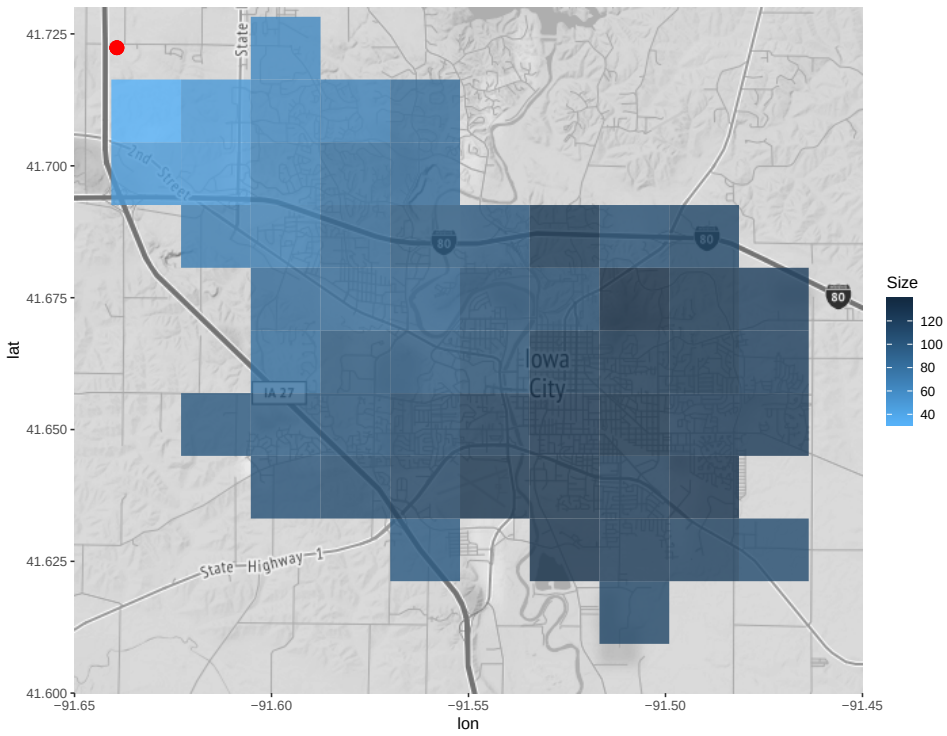

Figure 9 Average Time Window Size by Location

window sizes chosen by Algorithm 1, with lighter colors representing regions with smaller time windows, on average, and darker colors areas with larger time windows, on average.

The pattern in Figure 9 is clear: locations closest to the depot enjoy smaller time windows relative to locations further away. Because of their proximity to the depot, locations in the upper-left are typically sequenced toward the beginning or end of a service route. Thus, arrival at the start or finish of a workday is predictable and time windows are consequently narrower. In contrast, locations positioned beyond to/from trips to the depot experience more arrival time volatility. These customers often fall in the middle of a route and may appear earlier or later in the delivery sequence depending on where and when other service requests are realized. Because arrival times at these locations are less predictable, the assigned service time windows are therefore larger. We summarize the managerial implications for setting time window size based on location of request in Insight 5.

Insight 5 (Location of Request) *Assign larger time windows to customers further from the depot of operation and smaller time windows to closer customers.*

We conclude our discussion by exploring the joint impact of time and location of request on time window size. The heat map in Figure 10 illustrates the relationship, where the horizontal axis tracks time of request in minutes and the vertical axis measures travel distance from the depot in kilometers. The shade of each tile represents the average time window size selected by Algorithm 1 across all simulations with the specified time and location. Darker areas mark larger time windows and lighter areas mark smaller time windows. Unmarked tiles contain fewer than 10 observations and are omitted from the analysis.

Figure 10 largely confirms previous observations, that time window size decreases with time of request and increases with distance from the depot. The notable exception occurs in the bottom-left, where larger time windows are assigned to customers in close proximity to the depot who also place early requests. Such cases are illustrated by the center and right PDFs in Figure 3, where the probability masses are pushed to the beginning and end of the operating horizon. In these scenarios, because the time of request is early, the state does not contain enough information about other customer requests to reliably predict a position toward the beginning or end of the delivery sequence, hence the unusually large time window. However, as seen in the bottom row of tiles, customers with close proximity to the depot who make service requests later in the day receive smaller time windows. In these cases, position in the delivery sequence is more certain and the decrease in arrival time variability yields a smaller window for service.

6. Conclusion

We study the optimal setting of service time windows. Under mild assumptions on arrival time distributions, we show how to identify minimum expected-width time windows subject to a constraint on service level. Our analysis demonstrates that time windows should be set via a common density value across arrival time distributions. For the case of general distributional forms, our analytical results drive a heuristic that performs competitively relative to a dual bound and significantly outperforms benchmarks, including the industry standard. We glean five managerial insights. (i) The beginning and end of a time window should reflect the shape of the arrival time distribution. Consequently, the industry practice of assigning uniform fixed-width

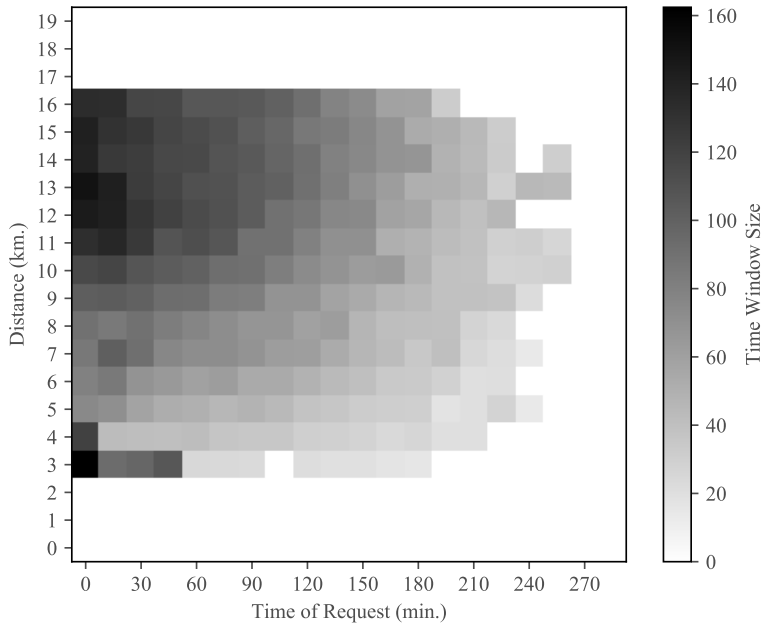

Figure 10 Average Time Window Size By Time of Request and Distance From Depot

time windows to all customers should be replaced by assignment of customer-specific time windows. (ii) Higher service levels require wider time windows. At a given service level, optimization is required to size time windows. (iii) Communicate two service time windows to customers whose arrival time distributions are bimodal. (iv) Assign larger time windows to earlier requests and smaller time windows to later requests. (v) Assign larger time windows to customers further from the depot of operation and smaller time windows to closer customers.

We suggest several directions for future research. As is still common in many service routing applications, our work assumes customer acceptance and routing decisions are exogenous to the process of setting time windows. Future work could consider the integration of these decisions, drawing on our research as a starting point. Further, Vareias et al. (2019) allows for customer-specific service levels. Future work could consider extending our results to that case. Additionally, we do not consider the possibility of revising time windows after they have been set. Real-time adjustments are already used in practice and are the subject of recent research (Dalmeijer et al. 2019). However, there does not seem to be academic work examining the setting of time windows at the time of the request with the possibility of future updates. Our results in this paper may

provide insight to such an analysis as well as a benchmark. Finally, future research might consider the role of pricing and incentives. For instance, while we focus on minimizing expected time window width to achieve a particular service level, one might also consider offering a customer a lower price to encourage the customer to choose a wider time window, potentially leveraging our analysis surrounding arrival time distributions. Larger windows would allow for more customers to be served on a given day without decreasing the firm's likelihood of meeting service commitments.

Acknowledgments

References

Agatz, Niels, Yingjie Fan, Daan Stam. 2021. The impact of green labels on time slot choice and operational sustainability. *Production and Operations Management* **30**(7) 2285–2303.

Apte, Aruna, Uday M Apte, Nandagopal Venugopal. 2007. Focusing on customer time in field service: A normative approach. *Production and Operations Management* **16**(2) 189–202.

Beskers, E. 2022. Time window assignment: A case study in retail distribution. Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands.

Campbell, Ann Melissa, Martin WP Savelsbergh. 2005. Decision support for consumer direct grocery initiatives. *Transportation Science* **39**(3) 313–327.

Dalmeijer, Kevin, Remy Spliet. 2018. A branch-and-cut algorithm for the time window assignment vehicle routing problem. *Computers & Operations Research* **89** 140–152.

Dalmeijer, Kevin, Remy Spliet, Albert Wagelmans. 2019. Dynamic time window adjustment. Tech. Rep. EI2019-22, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam. URL `http://hdl.handle.net/1765/116533`.

Ehmke, Jan F, Ann M Campbell. 2014. Customer acceptance mechanisms for home deliveries in metropolitan areas. *European Journal of Operational Research* **233**(1) 193–207.

Ellis, Blake. 2011. Waiting for the cable guy is costing us $38 billion. URL `http://money.cnn.com/2011/11/03/pf/cost_of_waiting/`. [Online; accessed 12-December-2022].

Flinterman, J. B. 2022. Assigning delivery time windows before all customers are known. Master's thesis, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, the Netherlands.

Gel, Esma S, Pinar Keskinocak, Tuba Yilmaz. 2020. Dynamic price and lead time quotation strategies to match demand and supply in make-to-order manufacturing environments. *Women in Industrial and Systems Engineering*. Springer, 541–560.

Hafızoğlu, A Baykal, Esma S Gel, Pınar Keskinocak. 2016. Price and lead time quotation for contract and spot customers. *Operations Research* **64**(2) 406–415.

Hermes. 2019. Standard service - enhancing your final mile solution. URL `https://www.hermesworld.com/en/our-services/distribution/parcel-delivery/services/standard-service.html`. [Online; accessed 10-September-2019].

Hoogeboom, Maaike, Yossiri Adulyasak, Wout Dullaert, Patrick Jaillet. 2021. The robust vehicle routing problem with time window assignments. *Transportation Science* **55**(2) 395–413.

Jabali, Ola, Roel Leus, Tom Van Woensel, Ton de Kok. 2015. Self-imposed time windows in vehicle routing problems. *OR Spectrum* **37**(2) 331–352.

Jalilvand, Mahdi, Mahdi Bashiri, Erfaneh Nikzad. 2021. An effective progressive hedging algorithm for the two-layers time window assignment vehicle routing problem in a stochastic environment. *Expert Systems with Applications* **165** 113877.

Kahvecioglu, Gökçe, Barış Balcıoglu. 2016. Coping with production time variability via dynamic lead-time quotation. *OR Spectrum* **38**(4) 877–898.

Keskinocak, Pinar, Sridhar Tayur. 2004. Due date management policies. D. Simchi-Levi, S.D. Wu, Z.J. Shen, eds., *Handbook of quantitative supply chain analysis*, *International Series in Operations Research & Management Science*, vol. 74. Springer, Boston, 485–554.

Köhler, Charlotte, Jan Fabian Ehmke, Ann Melissa Campbell. 2020. Flexible time window management for attended home deliveries. *Omega* **91** 102023.

Lim, Stanley Frederick WT, Qingchen Wang, Scott Webster. to appear. Do it right the first time: vehicle routing with home delivery attempt predictors. *Production and Operations Management* .

Madsen, Oli B. G., Kim Tosti, Jan Vælds. 1996. A heuristic method for dispatching repair men. *Annals of Operations Research* **61**(1) 213–226.

Martins, Sara, Manuel Ostermeier, Pedro Amorim, Alexander Hübner, Bernardo Almada-Lobo. 2019. Product-oriented time window assignment for a multi-compartment vehicle routing problem. *European Journal of Operational Research* **276**(3) 893–909.

Neves-Moreira, Fábio, Diogo Pereira Da Silva, Luís Guimarães, Pedro Amorim, Bernardo Almada-Lobo. 2018. The time window assignment vehicle routing problem with product dependent deliveries. *Transportation Research Part E: Logistics and Transportation Review* **116** 163–183.

Prisco, Jacopo. 2017. Why UPS trucks (almost) never turn left. URL `https://www.cnn.com/2017/02/16/world/ups-trucks-no-left-turns/index.html`. [Online; accessed 12-December-2022].

Spearman, Mark L, Rachel Q Zhang. 1999. Optimal lead time policies. *Management Science* **45**(2) 290–295.

Spliet, Remy, Said Dabia, Tom Van Woensel. 2018. The time window assignment vehicle routing problem with time-dependent travel times. *Transportation Science* **52**(2) 261–276.

Spliet, Remy, Guy Desaulniers. 2015. The discrete time window assignment vehicle routing problem. *European Journal of Operational Research* **244**(2) 379–391.

Spliet, Remy, Adriana F. Gabor. 2015. The time window assignment vehicle routing problem. *Transportation Science* **49**(4) 721–731.

Subramanyam, Anirudh, Akang Wang, Chrysanthos E Gounaris. 2018. A scenario decomposition algorithm for strategic time window assignment vehicle routing problems. *Transportation Research Part B: Methodological* **117** 296–317.

Ulmer, Marlin W, Barrett W Thomas. 2018. Enough waiting for the cable guy - estimating arrival times for service vehicle routing. *Transportation Science* **53**(3) 897–916.

Vareias, Anastasios D., Panagiotis P. Repoussis, Christos D. Tarantilis. 2019. Assessing customer service reliability in route planning with self-imposed time windows and stochastic travel times. *Transportation Science* **53**(1) 256–281.

Visser, Thomas, Martin Savelsbergh. 2019. Strategic time slot management: A priori routing for online grocery retailing. Tech. rep.

Waßmuth, Katrin, Charlotte Köhler, Niels Agatz, Moritz Fleischmann. 2022. Demand management for attended home delivery—a literature review. Technical note, available from http://dx.doi.org/10.2139/ssrn.405595, accessed on december 10, 2022, Erasmus Research Institute of Management.

Weiss, Darren. 2012. How Comcast killed the four-hour service window. URL https://fsd.servicemax.com/2012/01/18/how-comcast-killed-the-four-hour-service-window/. [Online; accessed 15-June-2020].

Yang, Xinan, Arne K Strauss, Christine SM Currie, Richard Eglese. 2014. Choice-based demand management and vehicle routing in e-fulfillment. *Transportation science* **50**(2) 473–488.

Yu, Xian, Siqian Shen, Babak Badri-Koohi, Haitham Seada. 2023. Time window optimization for attended home service delivery under multiple sources of uncertainties. *Computers & Operations Research* **150** 106045.

Zhang, Zhenzhen, Mengyang Liu, Andrew Lim. 2015. A memetic algorithm for the patient transportation problem. *Omega* **54** 60–71.

Otto von Guericke University Magdeburg
Faculty of Economics and Management
P.O. Box 4120 | 39016 Magdeburg | Germany

Tel.: +49 (0) 3 91 / 67-1 85 84
Fax: +49 (0) 3 91 / 67-1 21 20

www.fww.ovgu.de/femm

ISSN 1615-4274