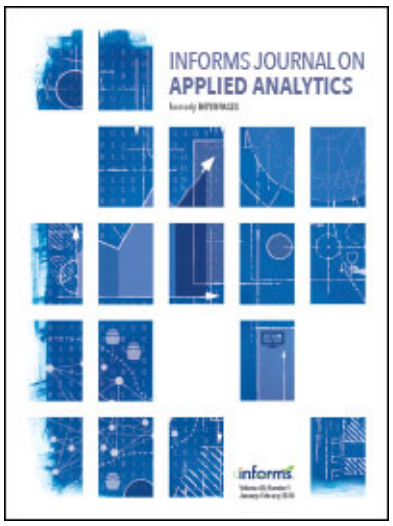


INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Redesigning Zoning Systems for Equitable and Efficient Last-Mile Delivery at Ninja Van

John Gunnar Carlsson, Stanley Frederick W. T. Lim, Sheng Liu, Han Yu, Witsanu Arntong, Ee Hsin Tan

To cite this article:
John Gunnar Carlsson, Stanley Frederick W. T. Lim, Sheng Liu, Han Yu, Witsanu Arntong, Ee Hsin Tan (2025) Redesigning Zoning Systems for Equitable and Efficient Last-Mile Delivery at Ninja Van. INFORMS Journal on Applied Analytics 55(5):412-423. https://doi.org/10.1287/inte.2025.0247

https://pubsonline.informs.org/journal/ijaa

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 55, No. 5, September–October 2025, pp. 412–423
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

Redesigning Zoning Systems for Equitable and Efficient Last-Mile Delivery at Ninja Van

John Gunnar Carlsson,[a] Stanley Frederick W. T. Lim,[b,*] Sheng Liu,[c] Han Yu,[a] Witsanu Arntong,[d] Ee Hsin Tan[d]

[a] University of Southern California, Los Angeles, California 90007; [b] Michigan State University, East Lansing, Michigan 48824; [c] University of Toronto, Toronto, Ontario M5S 1A1, Canada; [d] Ninja Logistics Pte. Ltd, Singapore 119967

*Corresponding author

Contact: jcarlsso@usc.edu, https://orcid.org/0000-0001-5346-8529 (JGC); slim@msu.edu, https://orcid.org/0000-0003-2211-9654 (SFWTL); sheng.liu@rotman.utoronto.ca, https://orcid.org/0000-0003-2365-6013 (SL); hyu376@usc.edu (HY); witsanu.a@ninjavan.co (WA); eehsin@ninjavan.co (EHT)

Received: January 13, 2025
Revised: April 17, 2025; June 12, 2025
Accepted: June 25, 2025

https://doi.org/10.1287/inte.2025.0247

Abstract. Efficient last-mile logistics is the key underpinning for sustainable e-commerce growth. As the final leg of delivery services, last-mile delivery is a time-consuming and labor-intensive process that requires careful operational planning and execution. In this paper, we describe the development of a novel zoning optimization framework that determines the assignment of customer locations to last-mile delivery stations for Ninja Van, a leading logistics service provider in Southeast Asia, to improve operational efficiency and work equity. The main technical development is a data-driven zoning optimization model that integrates the additively weighted Voronoi diagram and vehicle routing problem through a subgradient algorithm. The algorithm exploits the primal-dual formulation of the partitioning problem and is flexible enough to handle practical delivery scenarios with varying vehicle capacities and travel speeds. To the best of our knowledge, this is the first zoning optimization model that considers general multivehicle zones with uncertain demand, and these are the main contextual features of Ninja Van and many other logistics companies. The zoning algorithm we develop provides preferable theoretical guarantees. The implementation of the new zoning system at Ninja Van led to an average reduction of 6.6% in the work span for the delivery stations and a 3.5% reduction in driver delivery times. In addition to the monetary benefits from the shortened work hours, the new zoning system contributed to improved worker welfare by balancing workloads and limiting overtime. This zoning optimization also brought transparency and valuable insights into the management of delivery stations and drivers at Ninja Van.

History: This paper has been accepted for the *INFORMS Journal on Applied Analytics* Special Issue—2024 Daniel H. Wagner Prize for Excellence in the Practice of Advanced Analytics and Operations Research.
Funding: The research of S.F.W.T. Lim was supported by the Eli Broad College of Business 2024 Summer Research Grant.

Keywords: zoning • decision support • logistics • optimization

Introduction

The final leg of delivery service for parcels from a business to a consumer, also known as last-mile logistics, poses numerous challenges and substantial expenses for logistics firms, especially in developing nations. One major issue is meeting customer expectations for quick deliveries despite operational constraints, such as poor infrastructure, densely populated delivery areas, and unexpected disruptions (Lim 2023). The rapid growth of e-commerce and intense competition have resulted in higher expectations about the speed and reliability of deliveries. According to Statista (2024), e-commerce sales are projected to surpass $8 trillion by 2027, driven by the rise of online marketplaces and global internet usage. Fulfilling the rising expectations can become exceedingly costly with the last mile accounting for approximately 50% of total shipping costs (Lim 2024).

Additionally, there are widespread concerns about workload balance and fairness among last-mile delivery drivers. Uneven distribution of delivery tasks can cause inefficiencies, negatively affecting both delivery performance and driver morale. Particularly during peak periods, drivers often have to work overtime, raising the risk of workplace accidents and injuries (Maresca 2020).

To overcome these challenges, logistics companies resort to different investment and operational strategies. For example, a number of delivery service providers are embracing delivery management and route optimization software to better plan delivery routes. Alternative delivery options, such as curbside pickup, parcel lockers, and autonomous delivery systems, have also been considered. In addition to these technological initiatives, logistics companies, such as Singpost, J&T

Express, and Cainiao, are also seeking low-cost and easy-to-implement operational improvements, for example, the design of service zones.

Zoning is an effective strategy that carriers use to reduce lead times and improve delivery success rates. By dividing a large service area into smaller zones, each last-mile delivery station can serve its designated zone more effectively. This approach leverages drivers' local knowledge, simplifying resource distribution and management. However, optimally designing the zoning policy is a challenging problem because the practical region-level partitioning problem involves a large number of customer locations, and the direct mixed-integer programming formulation is computationally prohibitive. Consequently, in prior studies in the partitioning literature, researchers have developed specialized partitioning methods using parameterized diagrams, which tend to focus on single-vehicle zones and use the continuous approximation method to evaluate delivery workload. However, this approach is limited in its ability to model heterogeneous vehicle fleets (e.g., in terms of vehicle capacity) and more general routing objective functions. Despite challenges with current zoning methods, logistics companies have attempted to implement service zones using simple rules based on population density and past delivery data.

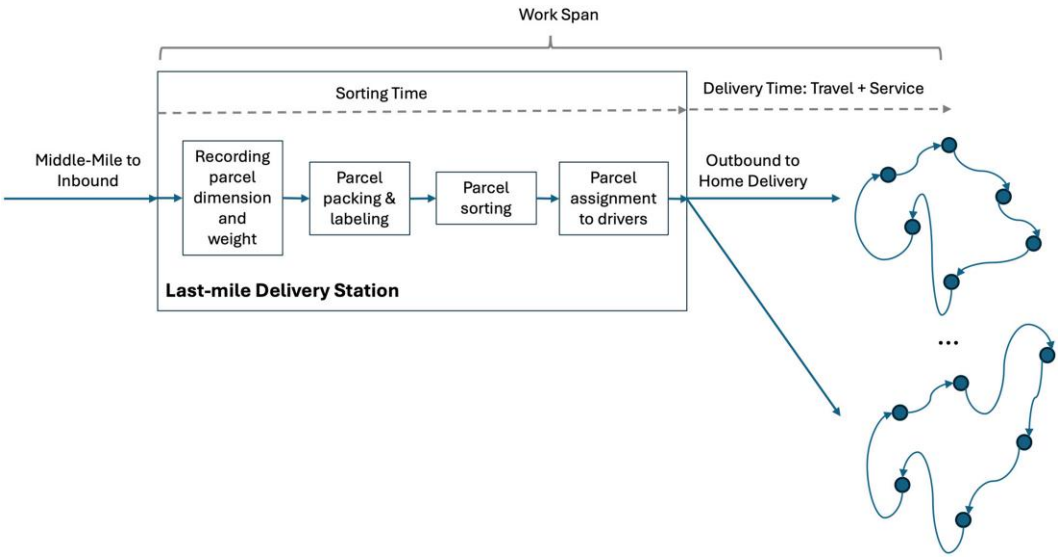
Ninja Van, which began operating in 2014, is a leading last-mile delivery company in Southeast Asia, offering extensive services across the region, including in Thailand, Malaysia, Vietnam, and Singapore. The company operates a network of delivery stations (depots), each with a fleet of vehicles and drivers, strategically located to optimize last-mile deliveries. This expansive network is essential to handle the large volume of deliveries, which averages one million parcels each day across Southeast Asia. Specifically, Ninja Van uses a zoning system that divides a large service area into smaller zones such that each last-mile delivery station is dedicated to serving the assigned zone. Figure 1 illustrates the inbound to the delivery station, the activities that take place within the station, and the outbound to customer homes. Under current zoning, the work span, defined as the duration between the start time of sorting the first parcel and the return time of the last driver upon finishing all delivery tasks assigned on the driver's route, is highly unbalanced across the stations. This discrepancy results in some stations having prolonged delivery times for customers and extended working hours for delivery drivers. This inequitable distribution also poses a significant risk for the company in maintaining its ongoing market expansion because increasing package volume may decrease customer service quality and the work conditions of the delivery drivers of its most vulnerable delivery stations.

In this paper, we discuss our collaboration with Ninja Van to develop a novel zoning optimization framework to design the optimal delivery zones (partitions) aimed at reducing work span and balancing workloads among zones and drivers. We implemented the proposed model on November 27, 2023, and evaluated the efficacy of the new algorithm through a pilot study in a major city in Southeast Asia.

Project Scope

The objective of this project is to improve the work span in last-mile delivery operations for Ninja Van. As defined earlier, the work span of a delivery station measures the length of the delivery process from sorting to

Figure 1. (Color online) The Parcel Delivery Process Flow



Note. ● denotes customer locations.

delivery. To improve the work span for all stations and achieve an equitable distribution of workload among stations and drivers, we propose to reduce the maximum work span of the delivery network. The direct and indirect implications of work span improvement to Ninja Van's stakeholders include

- Operational efficiency: Reducing the work span directly improves the efficiency of delivery operations. Shorter work spans mean faster delivery times, leading to higher throughput and better utilization of resources.
- Customer satisfaction: Meeting and exceeding customer expectations for fast and reliable deliveries is a key competitive advantage. Minimizing the work span ensures that parcels are delivered on time, improving customer satisfaction and loyalty.
- Service consistency: Reducing the work span gap among delivery stations enhances delivery service consistency in satisfying customers from different areas, thus contributing to a fair service provision.
- Driver well-being: Balanced workloads and shorter delivery routes contribute to better working conditions for drivers. Reducing overtime and ensuring fair task distribution can enhance driver morale and reduce turnover rates.

We focused on redesigning the zoning system to improve the work span, which involved reassigning delivery areas to delivery stations in Ninja Van's delivery network. The newly constructed delivery zones redefine the distribution of delivery jobs to delivery stations. The effective implementation of a zoning system is a critical operational pillar to allow Ninja Van to manage its resources and streamline the delivery process. To minimize the disruptions to the system and maintain the current labor contract agreement, we did not consider the options of changing the number of delivery stations or reassigning drivers among stations. In the "Technical Challenges in Zoning and Current Practice" section, we present how Ninja Van planned its zoning system and the technical challenges involved in the zoning system design.

Technical Challenges in Zoning and Current Practice

Ninja Van's current zoning system involves dividing the entire service region into smaller geographic zones. Each zone is strategically delineated based on several factors, including population density, geographic features, infrastructure availability, and historical delivery data. The company employs a heuristic approach based on these factors to determine the zone boundaries. Each zone is then assigned to a specific last-mile delivery station equipped with a fleet of vehicles and drivers. Figure 2 depicts the zones and their boundaries using Singapore as an illustration; each zone is served by a last-mile delivery station. Although this segmentation ensures that each subarea is a manageable size, it does not fully account for drivers' travel distances and future demand variability. This can result in prolonged travel times and work spans in some scenarios and significant disparities in delivery workload across zones. Although station managers and drivers receive similar financial compensation, some stations may handle much higher daily parcel volumes, and some drivers may need to exert more effort to complete their routes compared with others. For example, in the test city of our field trial, the average delivery time per driver was more than 10 hours for some stations (whereas many other stations observed five hours of average delivery time), and the station work span could be more than 20 hours under the current zoning. We compute an individual driver's delivery time per driver based on that driver's departure time from the delivery station for the first delivery and the return time to the same station after completion of the last delivery in the route. We then calculate the average delivery time per driver based on the average across all drivers' delivery time in our sample. Typically, zones were reviewed and manually updated when necessary, every six months to a year. Although the current process involves reconfiguring zones to better meet demand requirements, the process is laborious, time-consuming, and nonoptimal.

Figure 2. (Color online) Ninja Van's Current Zoning System Using an Example from Singapore

From a technical perspective, designing a good zoning system using data is a mathematically and computationally challenging task. Crucially, it requires a proper understanding of the relationship between zoning and work span, and this can be complicated because of the randomness of the data and the absence of an explicit form of the work span function. Therefore, Ninja Van relied on human judgment to draw a mapping from zoning to work span and execute the zoning design. Putting the zoning problem in the language of mathematical optimization, several technical

challenges can arise. First, we need to decide the assignment of each customer location (demand point) to a zone, and this implies that, assuming a continuous space, we have an infinite number of decision variables. Even if we restrict our attention to observed customer locations in one city, tens of thousands of locations would still require assignment by Ninja Van. Furthermore, the routing time of delivery vehicles is endogenous to zoning and the location of the station. This implies that evaluating the work span for a fixed zoning design has the same level of difficulty as solving a vehicle routing problem (VRP). As a result, a straightforward formulation of the zoning optimization problem that jointly determines customer location assignment and routing is computationally prohibitive for the practical instances that Ninja Van encountered.

Contributions to Theory and Practice

To solve the aforementioned challenges, we developed and implemented a new zoning system for Ninja Van, and it includes a number of unique features that advance the theory and practice of zoning optimization for last-mile logistics.

Embedding the VRP Solution in Zoning Optimization

Prior research on zoning and region partitioning in logistics primarily leverages continuous approximation techniques to design the optimal delivery zone (Ouyang 2007; Carlsson 2012; Carlsson et al. 2018; Banerjee et al. 2022, 2023). Although continuous approximation has merits in simplicity and tractability, it cannot fully capture practical delivery complexities, such as heterogeneous capacities or speeds and broader routing objectives such as work span. Notably, these studies mainly consider single-vehicle zones, and as a result, the partitioning process can be simplified (Ouyang 2007, Lei et al. 2012). In contrast, Ninja Van assigns multiple vehicles to each zone, and the job allocation among vehicles has a direct impact on the work span, and this requires new zoning ideas. To this end, we develop a flexible zoning optimization framework that explicitly integrates VRP solutions and allows us to model general multivehicle zones with diverse vehicle capacities. Doing so also opens the door for harnessing the power of state-of-the-art routing algorithms and heuristics in our zoning framework.

Data-Driven Solution Algorithm from Primal-Dual Formulation

The key building block of our zoning framework is a subgradient algorithm derived from the dual formulation of our zoning optimization model. Combining the convexity property of the dual problem and the structure of the work span function, we are able to establish the convergence condition of the subgradient algorithm. The condition is managerially relevant because it reflects how vehicle capacity impacts the work span. Moreover, because demand variability is significant, as observed from Ninja Van's operational data, we developed data-driven methods to evaluate the subgradient information using ideas from stochastic and robust optimization. This way, our solution algorithm naturally connects demand data with the final zoning outcome, which presents a unique data-driven zoning paradigm.

Difference-in-Differences (DiD) Analysis

To rigorously evaluate the efficacy of the proposed algorithm, rather than adopting the conventional practice of only comparing mean differences between the treatment group (i.e., stations that receive the new zones) and the control group (i.e., stations that continue using the current zones), we collected data before and after the implementation of the new zoning policy to facilitate a DiD experimental design. Through this approach, we can account not only for unobserved, time-invariant station and driver attributes but also for unobserved changes in conduct among competitors (e.g., competing couriers adjusting their service policies) caused by the new zoning policy. In our estimations, we use cluster-robust standard errors at the station and driver levels to account for contemporaneous dependence among deliveries at the individual station and driver, respectively. This approach tends to provide a more conservative estimate of statistical significance. We further strengthen our empirical identification by applying a propensity score weighting procedure to construct matched samples prior to implementing the DiD regressions.

Our Zoning Framework

We now introduce our zoning framework in four steps, from data collection to model validation.

Step 1: Data Collection

To facilitate model development, we collected a comprehensive set of operational data in the focal city from Ninja Van's database. This includes data on deliveries, route information, orders, station performance, and driver profiles. Table 1 summarizes the data sets and their respective variables. For example, the delivery data set includes order identification (ID), route ID, driver ID, delivery attempt date and time, customer location in terms of latitude and longitude, and outcome of the delivery attempt (i.e., a successful or a failed delivery). We merged these data sets to assemble the final data set for our analysis.

Table 1. Our Data Sets and Key Variables

Data set	Variables
Delivery data	Parcel ID, route ID, driver ID, attempt date time (e.g., 2023-11-05 15:23:18), customer location (latitude, longitude), attempt outcome (success, failure)
Route data	Route ID, driver ID, hub ID, route start time, route end time, total parcels for delivery, number of delivered parcels (i.e., successful delivery)
Order data	Order ID, order height, order length, order width, promised arrival time
Station data	Station ID, station location (latitude, longitude), number of employees, operating hours, sorting time
Driver data	Driver ID, employment type (full time, part time), employment start date, employment end date, vehicle type (2-wheel, 4-wheel)

Step 2: Input Calibration and Estimation

We needed to calibrate three parts of the work span function for the optimal zoning policy design: (i) sorting time at the station, (ii) en route travel time, and (iii) service time for on-site package delivery. For the sorting time, we trained a regression model using observed sorting time at each station as the target variable. The feature variables include the daily package volume, zone area size, number of zip codes, and sorting labor. The regression model estimates the sorting time as a function of the zoning decision. For en route travel time, we estimate the travel time between a pair of customer locations based on the ℓ_1 distance, which is adjusted to align with the practical road network. This is to save the use of commercial mapping tools, which Ninja Van did not employ and can be expensive to acquire for its daily use. We also consider different vehicle speeds for the two types of vehicles deployed by Ninja Van. For on-site service time, because there is a huge variability among drivers and reliably estimating a driver-specific and site-specific service time can be too noisy, we followed Ninja Van's suggestion to assume a five-minute fixed service time.

Step 3: Model Formulation and Solution Development for Zoning Optimization

We use K to denote the number of delivery stations. As we discuss above, Ninja Van seeks to minimize the maximum work span of delivery stations of its last-mile delivery network by reconfiguring the zoning system. This can be formulated as a general zoning optimization problem as follows:

Problem 1 (General Zoning Optimization for Last-Mile Delivery). The objective function is to minimize the maximum work span of delivery stations. Decision variables are the delivery zone for each delivery station $k = 1, \ldots, K$. Constraints are that delivery zones cover the whole service region and are nonoverlapping.

Specifically, we can define a work span function for each delivery station, and this depends on the delivery zone through the assigned customer locations and the location of the station. The decision variables are essentially geometric objects with boundaries and shapes that define the set of customer locations for which each delivery station is responsible. Because each zone is served by only one delivery station, the output from Problem 1 consists of K zones. Problem 1 aims to maneuver the shapes of delivery zones to redistribute delivery jobs among stations to ensure that the maximum work span is minimal. Note that the work span function generally does not admit a closed-form expression because the routing time hinges on the VRP solution, and the geometric objects cannot be described with a finite number of variables. The constraints of Problem 1 follow from the concept of Ninja Van's delivery system, in which each customer location in the service region must be assigned to exactly one delivery zone. Consequently, we can cast Problem 1 as a partitioning problem similar to political districting.

Addressing Problem 1 for Ninja Van poses two technical challenges. First, there are infinitely many ways to partition the service region into zones, and the shapes of zones can be arbitrary, which makes Problem 1 an infinite-dimension optimization problem that is difficult to solve or even evaluate. The intricate relationship between work span and the assigned demand locations only further complicates solving the problem. Second, observed package demand and customer locations vary from day to day; therefore, the work span function is random, and we need to consider its distribution when solving Problem 1. We discuss how our framework overcomes these two challenges below.

Restriction to a Special Class of Partitions. To obtain optimal zones in a tractable and interpretable way, we consider a class of partitions following the concept of an additively weighted Voronoi diagram (AWVD). Under an AWVD, a point $\mathbf{x}$ in the delivery region is assigned to a delivery zone for station k if $\mathbf{x}$ is closer to the station than any other stations measured by an additively weighted distance function. Specifically, the distance function is the difference between a typical distance function (e.g., ℓ_1 norm) and a station-specific weight parameter w_k: $dist(\mathbf{x}, k) - w_k$. The partition under

AWVD satisfies Equation (1):

$$\mathbf{x} \text{ is assigned to station } k \Leftrightarrow [dist(\mathbf{x}, k) - w_k] \leq [dist(\mathbf{x}, k') - w_k] \quad \forall k'. \tag{1}$$

Based on the above definition, when increasing the weight value of a station, the distance function value associated with the station tends to be lower, and as a result, more points are likely to be assigned to its delivery zone. We present example partitions from an AWVD in Figure 3, wherein the partition in Figure 3(b) is transformed from Figure 3(a) by increasing the weight of the station at the origin.

Given the simple definition from Equation (1), system operators can intuitively interpret the weight values and understand their relationships with the partition. Correspondingly, the zoning optimization problem is reduced to searching for the optimal weight values, $(W_1^*, \ldots, W_K^*)$ such that the optimal partition from the AWVD achieves the minimum work span. Admittedly, doing so does not examine all possible zoning compositions and may result in suboptimality. However, if we assume the work span function can be expressed as an integral of a measurable density function on the service region, an optimal AWVD, which balances the work span across stations and also minimizes the maximum work span (Carlsson et al. 2016), exists. Moreover, the optimal zones from the AWVD maintain contiguity and are easy to implement and manage. The contiguity property of the AWVD is particularly important to Ninja Van because the station managers would oppose delivery zones that are composed of disconnected areas.

Subgradient Algorithm. We solve for the optimal AWVD by a subgradient algorithm. To establish the validity of the subgradient algorithm, we consider a utility-maximizing equitable zoning problem with two inputs: (i) a station-specific utility function and (ii) a probability density function whose integral over the service region is normalized to one. We denote this density function by $f(\cdot)$.

Specifically, this problem takes the form of Problem 2.

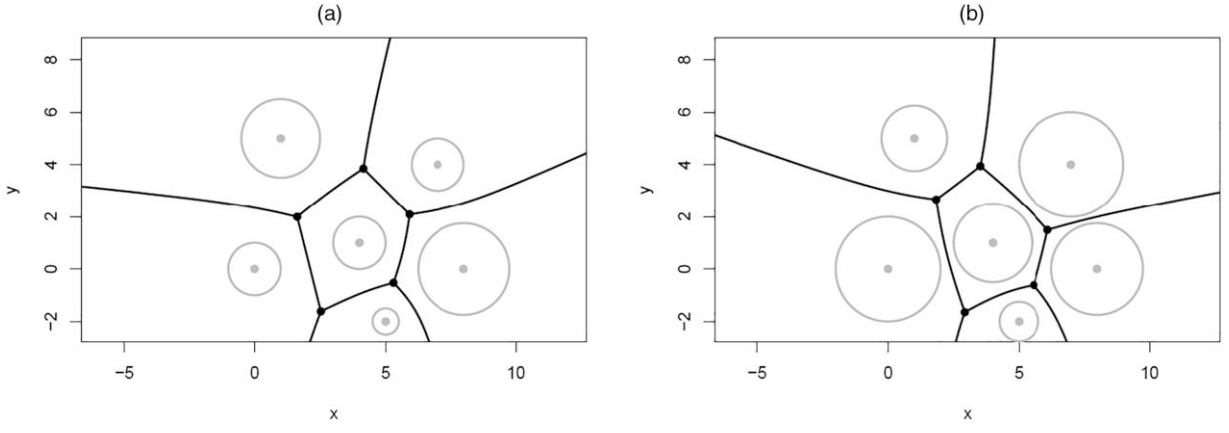
Problem 2 (Utility-Maximizing Equitable Zoning). The objective function is to maximize the aggregated utility over the service region. Decision variables are the zone assignment of every point in the service region. Constraints are that (i) the probability mass is equalized among zones and (ii) each point is assigned to exactly one zone.

In Problem 2, the goal is to maximize the overall utility of all zones by properly assigning points to zones, ensuring that each zone receives a fair amount of treatment according to the probability density function. The assignment decision is equivalent to the partition decision of Problem 1. Pavone et al. (2011) establish that the optimal solution to the dual of Problem 2 (the linear relaxation problem) gives an optimal AWVD that balances the probability mass among zones. Specifically, we can define the utility function to be negative of the distance function $dist(\mathbf{x}, k)$ for station $k = 1, \ldots, K$. Then, the optimal dual variables to constraints (i) of Problem 2 coincide with the desirable optimal weight values of the AWVD. Accordingly, the objective function of the dual problem is

$$\frac{1}{K}\sum_{k=1}^{K} w_i + \iint_{\text{service region}} f(\mathbf{x}) \max_k \left[-dist(\mathbf{x}, k) - w_k\right] d\mathbf{x},$$

where the dual variables are the weights of AWVD. Let $f(\cdot)$ be proportional to the work span density function if it exists; then, the derived AWVD equalizes the work span among zones.

Figure 3. Additively Weighted Voronoi Diagrams with Six Stations



Notes. Stations are the centers of the gray circles, and the circle radius indicates the weight of the AWVD. Partition (b) is derived from partition (a) by increasing the weight for station at point (0,0).

We can deduce from Carlsson et al. (2016) that a valid subgradient $G(w_1, \ldots, w_K) = (G_1, \ldots, G_K)$ to the convex dual problem satisfies

$$G_k = \frac{1}{K} - \text{normalized work span of zone } k, \quad (2)$$

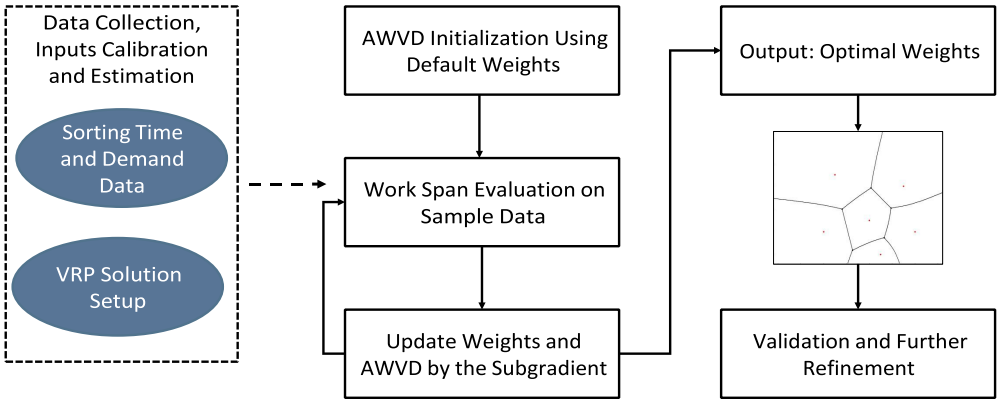
where the normalized work span can be interpreted as the ratio of the work span of zone k to the total work span of all zones; we provide the detailed derivation in the appendix. This implies that the subgradient is straightforward to obtain provided that we can evaluate the work span of each zone under the current partition of the AWVD. Specifically, we can solve the VRPs based on historical demand data given a partition. Running the subgradient algorithm based on Equation (2) iteratively allows us to arrive at the optimal weight value because of convexity.

The Convergence Conditions and Their Insights. As we discuss above, we are yet to formally establish the conditions under which the optimal AWVD can solve Problem 1 without any loss of optimality. The key is to show that the work span function can be expressed as an integral of an appropriately chosen $f(\cdot)$ (i.e., the work span density function). To this end, we examine the routing time that may not admit an explicit form for a VRP that minimizes the work span with heterogeneous vehicles. Ninja Van's current fleet consists mainly of small, two-wheeled vehicles that have limited capacity, which is complemented by large, full-size vans; therefore, we restrict our attention to these two types of vehicles and treat the full-size vans as loosely uncapacitated vehicles. It can be shown that the optimal work span is restricted by either (i) small, capacitated vehicles or (ii) large (uncapacitated) vehicles. In case (i), we can verify from the minimax principle that the optimal delivery job allocation is made so that each capacitated vehicle attains approximately the same work span as the uncapacitated vehicle. In this case, the optimal work span is approximately the optimal travel time from solving the traveling salesman problem (TSP) divided by the number of vehicles. Based on the Beardwood–Halton–Hammersley theorem, the optimal TSP tour length going through customer locations in an area can be estimated by the integral of $\sqrt{\rho(x)}$ over that area, where $\rho(\cdot)$ is the demand density function (Beardwood et al. 1959). Thus, the work span function admits an integral form, and the subgradient algorithm converges to the optimal solution for the AWVD. For case (ii), the previous argument works as long as the routing time of a large vehicle dominates, and the work span is mostly determined by its TSP tour length. Nevertheless, in general, the work span cannot be approximated as an integral function in case (ii) when the large vehicles take more time than the small ones. Companies such as Ninja Van that aim to compete with delivery speeds prioritize the use of small vehicles (often many) to reduce work span. Note that, although large vehicles carry more packages, they do not necessarily incur a longer routing time because they are often dispatched to areas with higher demand density (i.e., with shorter distances between locations), which is preferable for minimizing work spans. Therefore, we validate that the convergence condition for the proposed subgradient algorithm is not very restrictive with a simulation study in step 4 below.

Stochastic and Robust Solution. Applying the subgradient algorithm requires the estimation of the work span for each zone as indicated by Equation (2). As we mention above, the work span is random because of its dependence on stochastic demand. We implemented two approaches to evaluate the work span from historical demand data. The first approach follows the classic stochastic optimization routine to estimate the work span by sample average approximation (SAA), which draws samples from historical data and uses their average as the estimate. SAA tends to be more reliable because of its larger sample size. However, the computational burden of solving SAA increases substantially as the sample size increases as it does in our case, in which large-size VRPs must be solved for each sample. Alternatively, instead of increasing the sample size, we also consider a sample robust approach in which we take the maximum work span on the sample data as our estimate. This approach is more conservative than SAA because it plans for the worst case. To further accelerate the subgradient algorithm, we parallelize the computation of the VRP solution for each sample. We choose between stochastic and robust solutions based on validation using historical data. Figure 4 provides an illustrative process diagram of our zoning framework: the subgradient algorithm updates the weights of the AWVD iteratively, wherein each iteration involves evaluating the work span of every delivery zone based on VRP solutions on historical sorting time and demand data.

Advantages of Our Solution Approach. To summarize, our solution approach, which combines an AWVD and the subgradient algorithm, provides theoretical and practical benefits. From a technical standpoint, it has convergence guarantees under conditions that are relevant to Ninja Van and other delivery settings. This extends existing zoning research in logistics to handle multiple vehicles with different capacities and speeds serving the same zone. From a practical perspective, the zones obtained from the optimal AWVD are contiguous and easy to interpret. In particular, the partition is controlled by a single weight vector, and a manager can adjust the weight associated with a delivery station to account for other practical considerations in an intuitive

Figure 4. (Color online) Our Zoning Framework



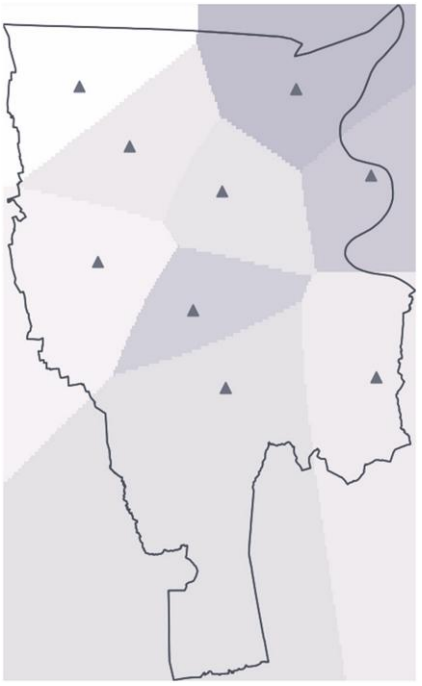
manner. For example, if a delivery station has a temporary capacity constraint because of a labor shortage, its weight can be increased to reduce the zone size and the package volume. Moreover, the subgradient algorithm is easy to implement and compatible with existing VRP algorithms: the normalized work span can be computed by calling any of the state-of-the-art VRP solvers, particularly open-source packages. This is particularly appealing for Ninja Van because it eases the need to develop in-house VRP algorithms, saving engineering costs and facilitating a fast implementation process. Finally, the compatibility of our approach with stochastic and robust optimization makes it a desirable choice for data-driven decision making. Depending on the data quality and underlying demand distribution, a manager can flexibly tweak our approach using data to improve out-of-sample performance.

Step 4: Validation Through Simulation

Using the calibrated inputs and parameters from step 2, we evaluate and validate the solution approach by simulating the zoning system obtained from one month of Ninja Van's data prior to implementation. First, we tested the efficiency of the subgradient algorithm on random data samples and found that the algorithm converged within 20 iterations. Given that the solution time for the VRP subroutines can be kept within minutes, the total solution time of our subgradient algorithm would only be a few hours for a reasonable number of samples. Such a solution-time requirement is relatively minor because the zoning problem does not need to be resolved often. We provide a sample zoning result from our algorithm in Ninja Van's focal city in Figure 5.

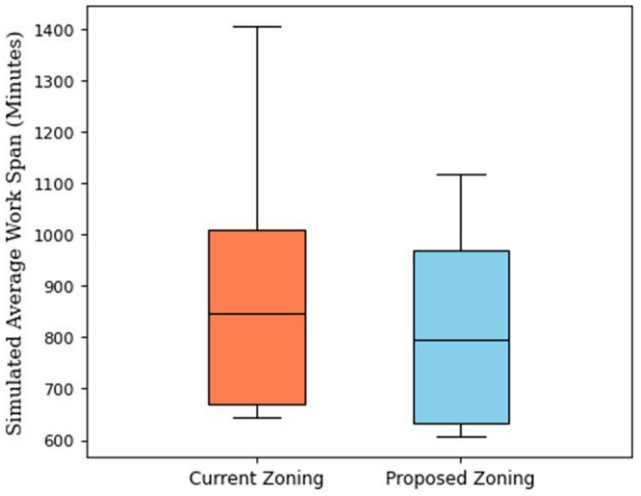
Then, we compare the zoning results from the SAA solution and the sample robust solution according to their out-of-sample work span performance, for which we split the data into a training set and a test set. For this specific month of data, the SAA solution leads to a slightly shorter maximum average time span than the sample robust solution on the test set, whereas both dominate the current zoning system. Specifically, we observe that our proposed zoning using SAA reduces the maximum average span by 20% compared with the current zoning. In terms of average driver delivery time, we observe a 17% reduction over the current zoning. We also verify that the proposed zoning induces a more equitable work span distribution: the gap between the maximum and minimum average work span is 33% lower than that of the current zoning, and the standard deviation of the work span decreases by 25%. We present the box plot of the station work span

Figure 5. (Color online) A Sample Zoning Result from the Proposed Zoning Optimization Model Using Ninja Van's Data



Note. Delivery stations are represented by triangles, each of which covers a shaded area (zone).

Figure 6. (Color online) Simulated Average Work Span Under Current and Proposed Zonings



from simulation in Figure 6, and this shows that the proposed zoning makes a notable change to the distribution of the work span by shifting it downward. Specifically, the median work span under the proposed zoning (right box plot) is lower than that of the current zoning (left box plot). Both groups exhibit similar interquartile ranges, suggesting that the variability across the middle 50% of stations remains comparable. However, the proposed zoning has a shorter overall range—from minimum to maximum—indicating reduced extremes in workload and a more consistent distribution of work across stations.

Implementation

Supported by the promising simulation result, we implemented the proposed zoning system in a major city in Southeast Asia after deliberating with Ninja Van on the zoning result. The implementation was set up to facilitate a DiD analysis to estimate the causal effect of the new zoning system.

Ninja Van Inputs to Refine the Model

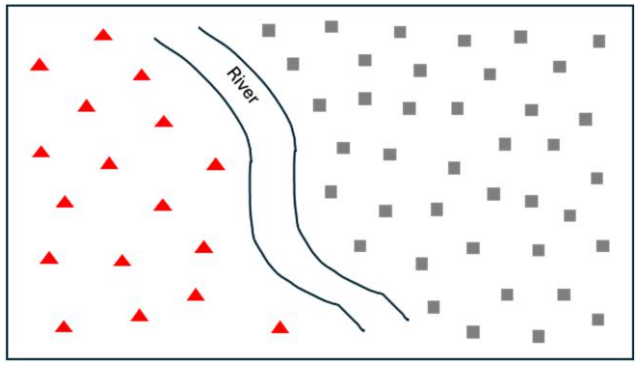
We discussed the proposed zoning system with Ninja Van's senior management team before the zoning was scheduled to update. In addition to the key metrics, such as average work span and (driver) delivery time, we also presented the 90% quantile of delivery time, probability of meeting the 12-hour work span threshold, average station order volume, and average driver order volume. These data provided additional insights into the distribution of workload among stations and drivers. Ninja Van's operations team reviewed the results and proposed several adjustments using on-the-ground operational constraints and labor considerations. For example, the zoning was revised to conform with geographical boundaries, such as main roads and rivers. We also made adjustments to ensure that the order volume of delivery stations and the number of drivers did not drop below a specific threshold to maintain operational efficiency and workforce engagement. Further discussions were held to fine-tune the zoning plans, taking into account potential impacts on service quality and customer satisfaction. The management team approved the final revised zoning plan, which was scheduled for implementation on November 27, 2023. This collaborative approach not only ensured that we were able to rigorously evaluate the efficacy of the proposed zoning system but also that it was viable from an operational perspective and aligned with the company's strategic goals.

Field Implementation

We implemented the pilot in a major city in Southeast Asia, in which the test area is separated by a river. We randomly selected the area (and corresponding stations) on the left side of the river as the treatment area and stations to receive the new zones and the area (and corresponding stations) on the right side of the river as a control area in which stations receive no adjustments to their current zoning policy. This setup allowed us to control for environmental and market factors that affected our results. The treatment area included 17 stations and 369 drivers, whereas the control area comprised 40 stations and 850 drivers (see Figure 7). Comparing the results before and after the implementation with both the treatment and control areas provided us with rigorous estimates of the benefits of the proposed zoning system (and also the statistical significance of the results).

Ninja Van launched the new zoning system on November 27, 2023. To establish a baseline, the company avoided making any significant operational adjustments or policy changes from October 26 to November 26. Each delivery station fell under the

Figure 7. (Color online) The Trial Area



Notes. ▲ and ■ denote treatment and control stations, respectively. The diagram is not drawn to scale and the station locations do not represent the locations we implemented in practice.

responsibility of a station manager. All station managers in the treatment area implemented the new zones effective November 27. The drivers in these stations were not informed of any zoning changes; they continued to receive their daily package assignments from their station managers. We monitored all deliveries until approximately December 20 to avoid monitoring during the Christmas period.

Benefits and Impact

In this section, we discuss three aspects of the benefits and impact: improved time performance (i.e., nonmonetary), monetary benefits, and managerial takeaways.

Improved Time Performance

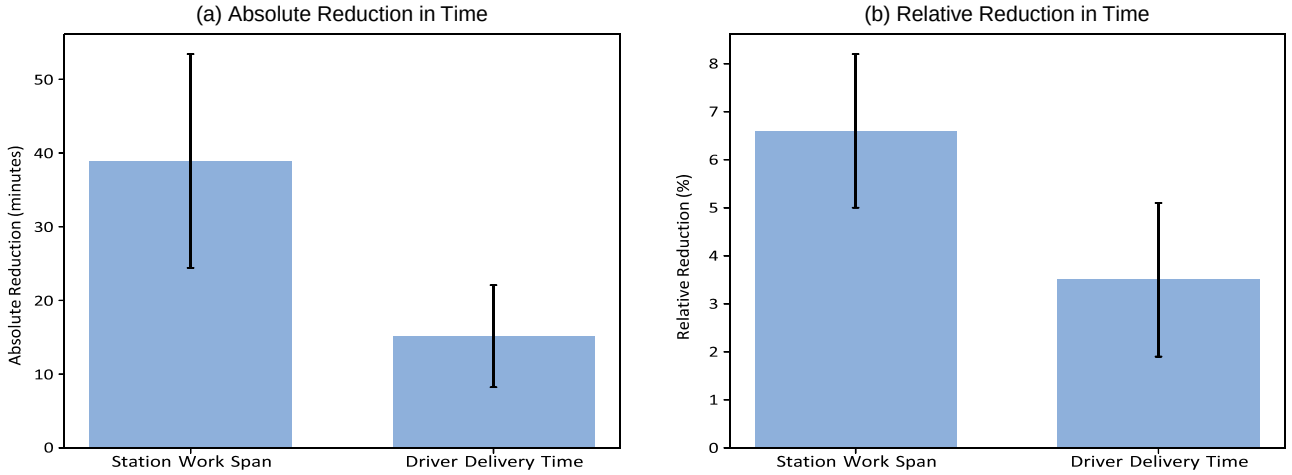
Consistent with the objective function of our zoning, the direct benefit of the zoning system we developed lies in the reduced work span. Based on the field data, we documented that the average work span for stations in the treatment area is 6.6% (i.e., 38.92 minutes) shorter than that of stations in the control area (see Figure 8(a)). We achieved this without affecting the total demand and driver allocation. Because the work span is closely related to the delivery time of drivers, we also observed that the average delivery time per driver was reduced by 3.5% (i.e., 15.15 minutes) for drivers on routes in the treatment area compared with those in the control area (see Figure 8(b)). Naturally, the decrease in delivery times could boost customer satisfaction as a result of the improved delivery service quality. In addition, reduced delivery time implies that the drivers could finish their work earlier, limiting their overtime work.

We also assessed the balancing benefit of the new zoning system. Specifically, we documented that the implemented system led to a 6.7% lower standard deviation of station work span and a 13.6% lower standard deviation of driver delivery time. These were consistent with the simulation and our expectations, but we note that we did not observe the differences in standard deviation at the 0.05 significance level because of a larger standard error than that of the mean effect. According to our analysis and Ninja Van's practices, we attribute the lower statistical significance to two possible reasons. First, in actual practice, drivers may not adhere strictly to the optimal route sequences identified in our numerical analysis. Previous research documents the problem of drivers deviating from system-recommended routes (Liu et al. 2021, Lim et al. 2023). Second, upon the focal company's request, we imposed additional constraints on the final zoning plan to guarantee a minimum average driver volume per station. This constraint was intended to ensure each driver handles a minimum volume daily, but it inherently led to deviations from the optimal assignments. We anticipate seeing more significant improvements in the key metrics if these constraints are not in place. In particular, Ninja Van would observe shorter and more balanced delivery times among its drivers as it moves to improve the routing algorithm and drivers' adherence, and this can form a positive feedback loop to the zoning system we developed.

Monetary Benefits

We assess the economic impact of our zoning algorithm on the costs of operating the delivery stations in the focal market. Using monthly operating costs, which include fixed (e.g., lease equipment, miscellaneous) and variable (e.g., labor, utilities, security, cleaning) expenses, we estimate the average cost per station is 82,353.38 Thai Baht (or US$2,223.54). About 50% of this cost is variable. Our model suggests a 6.6% reduction in these variable costs. With 57 delivery stations, this means a monthly savings of 2,717.66 Thai Baht (US$89.80) per station, totaling 1,858,879.44 Thai Baht (US$50,189.74) annually for the focal market. Nationwide, with 500

Figure 8. (Color online) The Improvement of the Proposed Zoning vs. Current Zoning in the Field Experiment



stations, we project the annual savings will reach 16,305,960 Thai Baht (US$440,260.92).

Managerial Takeaways

The new zoning system provided the management team at Ninja Van full transparency into how customers from different areas were assigned to delivery stations. It minimized subjective biases from human judgment and facilitated automating the zoning decision making in a data-driven way. More importantly, learning the mapping from zoning to work span distribution enabled Ninja Van to identify the bottleneck stations at which the staffing decision could be improved. For Ninja Van and many logistics companies in developing countries, disruptive technologies such as drones and automated sorting systems are an appealing but costly investment. The new zoning system presented a scalable and low-cost solution that could lead to improved operational outcomes without making sweeping changes to the existing infrastructure and labor agreements.

Long-Term Implications

The implementation of the new zoning solution has several long-term implications for Ninja Van's operational efficiency and agility, customer satisfaction, and worker welfare. First, by strategically adjusting delivery zones, companies such as Ninja Van can optimize delivery routes, reduce fuel consumption, and lower vehicle maintenance costs. This efficient allocation of resources not only cuts down on operational expenses but also enhances overall productivity. Maintaining a high level of productivity is critical to Ninja Van as it looks to sustain growing delivery demand in Southeast Asia. Moreover, the zoning system we develop is sufficiently flexible to reflect different objective functions and workforce/vehicle compositions, which will allow Ninja Van to quickly adjust to different market conditions. For example, if crowd-sourcing drivers is introduced, the zoning system can be updated by simply revising the VRP solution subroutine to adjust for uncertain driver schedules. Second, optimized delivery zones enable more accurate and consistent delivery times, better meeting customer expectations. Furthermore, faster and more reliable deliveries imply higher customer satisfaction and loyalty, and these are essential for Ninja Van's competitive advantage. The data-driven zoning approach also allows for better service differentiation, catering to specific regional demands and preferences. Finally, proper zoning and workload distribution help prevent overworked drivers, reducing the risk of accidents and improving overall job satisfaction. By limiting long working hours and balancing workloads, the company can enhance worker welfare, and this, in turn, will lead to lower turnover rates and higher employee retention. This holistic approach not only benefits the workforce but also enhances the company's reputation as a responsible employer, creating a win–win scenario for both the company and its drivers.

Summary

In this study, we describe our development of a new data-driven delivery zoning system that Ninja Van implemented to improve its last-mile delivery operations. Our field implementation documented reduced work spans, delivery times, and operational cost savings that, by extrapolation from monthly to yearly estimates, amount to roughly half a million U.S. dollars per year. Specifically, we present an easy-to-implement and flexible four-step zoning optimization framework. The technical core of our framework (step 3: model formulation and solution development for zoning optimization) involves iteratively solving VRPs using actual demand data and dividing the delivery area with additively weighted Voronoi diagrams. By doing so, we are the first to integrate the VRP solution framework with additively weighted Voronoi diagrams for use in partitioning applications that account for demand uncertainty. The key recipe for deriving solutions efficiently is to leverage the easily computable subgradient information from the dual formulation of the partitioning problem. The computation of the subgradient is adaptable to stochastic and robust optimization formulations and compatible with state-of-the-art routing algorithms. This framework offers a structured approach for other logistics companies to examine alternative service zoning policies and direct their zoning processes to be data-driven and oriented to fairness. We also expect the zoning framework to be applicable to nonprofit settings, such as humanitarian logistics, in which response time represents a key objective.

Acknowledgments

The authors express sincere gratitude to Ninja Van and its operations team for their invaluable contributions to this research. The authors are also grateful to seminar attendees at the Eindhoven University of Technology and the final presentation session of the 2024 Wagner Prize competition for their comments.

Appendix. Optimization Model Formulation and Solution

We present first the general zoning optimization model with deterministic demand $\mathbf{d}$. Specifically, let $\mathbf{d} \in \mathbb{Z}^M$ be a demand vector for M possible demand locations, where d_m is the order quantity from location m. The decision variables are denoted by $\{R_1, \ldots, R_K\}$, where R_k represents zone k and is a subset of the service region. Furthermore, let $\text{Workspan}(\mathbf{d}, p_k, R_k)$ denote the work span function for station k, which depends on the depot location p_k and

zone R_k. Then, Problem 1 can be formulated as

$$\min_{R_1,\dots,R_K} \max_{1\le k\le K} \text{Workspan}(\mathbf{d}, p_k, R_k), \tag{A.1}$$

$$s.t. \quad \bigcup_{k=1}^{K} R_k = \text{service region}, \tag{A.2}$$

$$R_k \cap R_{k'} = 0, \ \forall k \ne k' \ k,k' = 1,\dots,K, \tag{A.3}$$

where Constraint (A.2) ensures the coverage of the service region and Constraints (A.3) ensure that the zones are nonoverlapping. In practice, the observed order demand varies from day to day, and so does the work span of every station. Following the SAA scheme, we can replace Workspan $(\mathbf{d}, p_k, R_k)$ by the sample average $\sum_{n=1}^{N} \text{Workspan}(\mathbf{d}_n, p_k, R_k)/N$ over N demand samples $\{\mathbf{d}_1,\dots,\mathbf{d}_N\}$. Accordingly, the sample average work span is used for gradient computation in Equation (2). Similarly, we can use the sample worst case work span in the case of robust optimization.

We then present more details about the dual problem for finding the optimal AWVD, which is also discussed in Carlsson et al. (2016). Recall that the dual problem is

$$\min_{\mathbf{w}} \frac{1}{K}\sum_{k=1}^{K} w_i + \iint_{\text{service region}} f(\mathbf{x}) \max_k [-dist(\mathbf{x},k) - w_k] d\mathbf{x},$$

and we can verify that a subgradient of the above objective function $G(w_1,\dots,w_K) = (G_1,\dots,G_K)$ can be computed by

$$G_k = \frac{1}{K} - \iint_{R_k(\mathbf{w})} f(\mathbf{x}) d\mathbf{x}, \ k = 1,\dots,K.$$

This is because (let $G^{\Delta} = (G_1^{\Delta},\dots,G_K^{\Delta})$ with $G_k^{\Delta} = -\iint_{R_k(\mathbf{w})} f(\mathbf{x})d\mathbf{x}$) for any $\mathbf{w}$ and $\mathbf{w}'$, the following inequality holds:

$$\begin{aligned}
&\iint_{\text{service region}} f(\mathbf{x}) \max_k [-dist(\mathbf{x},k) - w_k'] d\mathbf{x} \\
&\ge \sum_{k=1}^{K} \left\{ \iint_{R_k(\mathbf{w})} f(\mathbf{x})(-dist(\mathbf{x},k) - w_k) d\mathbf{x} + G_k^{\Delta}(w_k' - w_k) \right\} \\
&= \iint_{\text{service region}} f(\mathbf{x}) \max_k [-dist(\mathbf{x},k) - w_k] d\mathbf{x} + G^{\Delta}(\mathbf{w}' - \mathbf{w}).
\end{aligned}$$

With historical data, we can evaluate $\iint_{R_k(\mathbf{w})} f(\mathbf{x})d\mathbf{x}$ as the normalized work span of zone k.

References

Banerjee D, Erera AL, Toriello A (2022) Fleet sizing and service region partitioning for same-day delivery systems. *Transportation Sci.* 56(5):1327–1347.

Banerjee D, Erera AL, Stroh AM, Toriello A (2023) Who has access to e-commerce and when? Time-varying service regions in same-day delivery. *Transportation Res. Part B Methodological* 170:148–168.

Beardwood J, Halton JH, Hammersley JM (1959) The shortest path through many points. *Math. Proc. Cambridge Philos. Soc.* 55(4):299–327.

Carlsson JG (2012) Dividing a territory among several vehicles. *INFORMS J. Comput.* 24(4):565–577.

Carlsson JG, Behroozi M, Mihic K (2018) Wasserstein distance and the distributionally robust TSP. *Oper. Res.* 66(6):1603–1624.

Carlsson JG, Carlsson E, Devulapalli R (2016) Shadow prices in territory division. *Networks Spatial Econom.* 16(3):893–931.

Lei H, Laporte G, Guo B (2012) Districting for routing with stochastic customers. *EURO J. Transportation Logist.* 1(1–2):67–85.

Lim SFWT (2023) Why so many packages don't get delivered. Harvard Business Review. Accessed August 25, 2024, https://hbr.org/2023/11/research-why-so-many-packages-dont-get-delivered.

Lim SFWT (2024) Cutting last-mile delivery costs: New tactics can boost profitability and satisfaction with subscription services. *MIT Sloan Management Rev.* 66(2):15–17.

Lim SFWT, Wang Q, Webster S (2023) Do it right the first time: Vehicle routing with home delivery attempt predictors. *Production Oper. Management* 32(4):1262–1284.

Liu S, He L, Max Shen ZJ (2021) On-time last-mile delivery: Order assignment with travel-time predictors. *Management Sci.* 67(7):4095–4119.

Maresca T (2020) Delivery drivers working to death amid online shopping boom in S. Korea. UPI. Accessed December 14, 2023, https://www.upi.com/Top_News/World-News/2020/12/24/Delivery-drivers-working-to-death-amid-online-shopping-boom-in-S-Korea/1141608844270/.

Ouyang Y (2007) Design of vehicle routing zones for large-scale distribution systems. *Transportation Res. Part B Methodological* 41(10):1079–1093.

Pavone M, Arsie A, Frazzoli E, Bullo F (2011) Distributed algorithms for environment partitioning in mobile robotic networks. *IEEE Trans. Automatic Control* 56(8):1834–1848.

Statista (2024) Retail e-commerce sales worldwide from 2014 to 2027. Accessed June 13, 2024, https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

John Gunnar Carlsson is the Kellner Family Associate Professor of Industrial and Systems Engineering at the University of Southern California. He serves as an associate editor for *Operations Research*, *Management Science*, and *Transportation Science* and his research is concerned with applications of computational geometry and probability theory to problems in transportation and logistics. He is the recipient of the Transportation Science and Logistics Society Best Paper Award (Freight Special Interest Group).

Stanley Frederick W. T. Lim is an associate professor of supply chain management at Michigan State University, where he codirects the food access and supply chain technology lab. His research focuses on the economics of distribution services and solving operational problems in downstream or so-called last-mile supply chains, spanning both digital and nondigital retail contexts. He has collaborated with retailers, logistics providers, and nonprofits across the Americas, Europe, and Asia.

Sheng Liu is an assistant professor of operations management and statistics at the Rotman School of Management, University of Toronto. His research focuses on solving operations problems in supply chains, transportation, and logistics systems through optimization and data analytics. He has collaborated with large retailers, including JD.com and Sport Chek.

Han Yu is a software engineer at Google, specializing in agent quality and agent modality in artificial intelligence. She earned her PhD in industrial and system engineering from the University of Southern California in 2023, where her research focused on last-mile delivery systems, time series transportation forecasting, dispatching systems, and network optimization.

Witsanu Arntong oversees operations at Ninja Van Thailand, where he provided support to this research. He implemented and tested the study's recommendations and assumptions within a real-world case study, demonstrating the practical applicability of its findings and validating their effectiveness within complex operational environments, particularly for delivery challenges.

Ee Hsin Tan is a network planner with the regional team at Ninja Van, where she streamlines operations and supports scalable growth across Southeast Asia. Her work includes driving the deployment of automated sorting systems in warehouses and enabling data-driven scheduling for linehaul to enhance network efficiency.