Equitable Delivery Zoning for Last-Mile Logistics: A Framework Validated with Implementation

John Gunnar Carlsson[a], Stanley Lim[b], Sheng Liu[c,*], Han Yu[a]

[a] Daniel J. Epstein Department of Industrial & Systems Engineering, University of Southern California, USA

[b] Eli Broad College of Business, Michigan State University, USA

[c] Rotman School of Management, University of Toronto, Canada

[*] Corresponding author

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

March 4, 2025

Keywords: Last-mile logistics, parcel delivery, zoning optimization, Voronoi diagrams

1 Introduction

Last-mile logistics, particularly in developing countries, is fraught with challenges and significant costs. One of the main challenges is meeting customer expectations for rapid delivery amid constraints such as poor infrastructure and high delivery densities. The last mile accounts for a substantial portion (41-53%) of the total supply chain costs, attributed to factors such as low route density, traffic congestion, and high labor expenses (Sykes, 2023). Workload imbalance and fairness concerns among drivers are also significant, as unequal distribution of delivery tasks can lead to inefficiencies, impacting overall delivery performance and driver morale. In particular, overtime work becomes common among delivery drivers during busy seasons, causing injuries or even deaths on the job (United Press International, 2020).

Delivery companies often make decisions on how to assign delivery stations, vehicles, and drivers to different service areas to meet customer expectations for fast delivery. They generally have two options: either maintain one large service area that is served by all assets together, or divide the service area into smaller sub-areas or zones (Zhong *et al.*, 2007). The second strategy allows companies to take advantage of the drivers' familiarity with their local zones, leading to better delivery performance in terms of first-attempt success rates and on-time delivery. This approach also helps drivers be more productive and deliver faster because they have a smaller area to cover. However, optimally designing the zoning policy is a challenging problem. This is 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, prior studies in the partitioning literature 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 (Ouyang, 2007).

In this paper, we collaborate with our industry partner, Ninja Van[1], to develop a zoning optimization model that determines optimal zones by prioritizing delivery lead time reduction and workload balance. Specifically, the model seeks to minimize the maximum work span while reducing disparity in workload among delivery stations and drivers. We define the delivery work span of a delivery station 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 their route.* In

[1] Ninja Van is Southeast Asia's leading logistics provider, with the highest service coverage over 6 countries in the region.

turn, the maximum work span measures the longest work span observed in the delivery network within a given time period. We solve the zoning optimization model by leveraging the primal-dual properties of Voronoi diagrams that minimize the expected or the worst-case work span of depots. To the best of our knowledge, this is the first zoning optimization model that considers general multi-vehicle zones with uncertain demand, which are the main contextual features of Ninja Van and many other logistics companies. The proposed zoning framework was implemented and validated by Ninja Van.

2 Model and Algorithm

We consider a service region C with n depots $\{p_1, \ldots, p_n\}$. Depot p_i serves as a last-mile delivery station for service zone R_i. Each depot is responsible for dispatching a group of drivers who deliver packages to customers in the associated service zone. We use V_i to denote the set of drivers servicing depot i. Each driver has limited delivery capacity, which depends on the type of vehicle used by the driver. In particular, the focal company deploys two types of vehicle: two-wheeled vehicles and full-size delivery vans. Two-wheeled vehicles are less costly and more flexible to manage, whereas full-size vans can deliver many more packages per trip. Because the driver hiring decision is administered by the station managers, the driver cannot be reassigned easily. Therefore, we focus on the service zoning decision, with the zones denoted by $\{R_1, R_2, \ldots, R_n\}$, while fixing the size of the fleet at each depot. Based on the service requirement and the nonoverlapping condition, the service zones must satisfy $\bigcup_{i=1}^{n} R_i = C$ and $R_i \cap R_j = 0$ for $i \neq j$ and $i, j = 1, \ldots, n$.

Let $F(\mathbf{d}, p_i, R_i)$ be the delivery performance metric (e.g., time or cost) of depot p_i, which depends on the zoning through R_i. Our model does not limit the form of $F(\mathbf{d}, p_i, R_i)$ but it is natural to assume $F(\mathbf{d}, p_i, R_i) \leq F(\mathbf{d}, p_i, R_j)$ if $R_i \subseteq R_j$ and $F(\mathbf{d}, p_i, R_i) \leq F(\mathbf{d}', p_i, R_j)$ if $\mathbf{d} \leq \mathbf{d}'$ (monotonicity in zone and demand). In the remainder of this paper, we take $F(\mathbf{d}, p_i, R_i)$ to be the delivery work span of depot p_i, that is, *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 their route.* In Ninja Van's case, $F(\mathbf{d}, p_i, R_i)$ includes the sorting time of packages, travel time along the road network, and on-site service time to drop off packages. Note that $F(\mathbf{d}, p_i, R_i)$ itself is the optimal objective function of a routing optimization problem because the travel time and on-site service time of drivers depend on the assigned routes (while sorting time can be expressed as a function of the volume of orders at the depot). Mathematically, the work span function can be written as

$$S(\mathbf{d}', p_i, R_i) + \max_{v \in V_i} \; RT_v(\mathbf{d}', p_i, R_i), \; \forall i = 1, \ldots, n, \tag{1}$$

where $S(\mathbf{d}', p_i, R_i)$ represents the total sorting time at depot p_i and $RT_v(\mathbf{d}', p_i, R_i)$ is the total routing time of driver v at depot p_i, the latter of which includes both travel time and on-site service time. Note that $RT_v(\mathbf{d}', p_i, R_i)$ comes from the VRP solution that decides the optimal routes of drivers given the demand information.

The objective of the delivery zoning problem is to minimize the maximum work span of depots. Doing so shortens the maximum delivery time to customers and increases the same-day delivery success rate (a longer delivery time increases the chance of rescheduling a delivery job), which are the business priorities of Ninja Van. The minimax objective also encourages a more equitable distribution of the workload among delivery stations and drivers. The corresponding zoning optimization problem can be formulated as

$$\min_{R_1, \ldots, R_n} \max_{1 \leq i \leq n} F(\mathbf{d}, p_i, R_i) \tag{2}$$

$$\text{subject to } \bigcup_{i=1}^{n} R_i = C, \tag{3}$$

$$R_i \cap R_j = 0, \; \forall i \neq j \;\; i, j = 1, \ldots, n., \tag{4}$$

2.1 Solution Algorithm

To obtain optimal zones in a tractable and interpretable way, we consider a class of partitions following the concept of *additively weighted Voronoi diagram* (AWVD). Under AWVD, a point $\mathbf{x}$ in the delivery region is assigned to a delivery zone for station k if $\mathbf{x}$ is closest 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

$$\mathbf{x} \text{ is assigned to station } k \Leftrightarrow dist(\mathbf{x}, k) - w_k \leq dist(\mathbf{x}, k') - w_{k'} \quad \forall k'. \tag{5}$$

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. Given the simple definition from Equation (5), 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 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 that balances the work span across stations exists, which also minimizes the maximum work span (Carlsson *et al.*, 2016). Moreover, the optimal zones from AWVD maintain contiguity and are easy to implement and manage. The contiguity property of AWVD is particularly important to Ninja Van because the station managers would oppose delivery zones that are composed of disconnected areas.

We solve for the optimal AWVD using a subgradient algorithm. Specifically, we leverage the primal-dual relationship between the AWVD partition problem and a utility maximization problem and derive the subgradient $G(w_1, \ldots, w_K) = (G_1, \ldots, G_K)$ to the convex dual problem as (Pavone *et al.*, 2011, Carlsson *et al.*, 2016):

$$G_k = \frac{1}{K} - \text{normalized work span of zone } k, \tag{6}$$

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. This implies that the subgradient is rather straightforward to obtain so long as we can evaluate the work span of each zone under the current partition of AWVD. Running the subgradient algorithm based on formula (6) iteratively would arrive at the optimal weight value due to convexity.

2.2 Convergence Condition

We demonstrate that the subgradient algorithm converges under mild conditions. 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 problem that minimizes the work span with heterogeneous vehicles. Ninja Van's current fleet mainly consists of small two-wheeled vehicles that have very limited capacity, which is complemented by large full-size vans, so 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 either restricted by (a) small capacitated vehicles or (b) large (uncapacitated) vehicles. In case (a), 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 Bearwood–Halton–Hammersley Theorem, the optimal TSP tour length going through customer locations in an area can be estimated by the integral of $\sqrt{\rho(\mathbf{x})}$ 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 AWVD. For case (b), 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 can not be approximated as an integral function in case (b) when the large vehicles take more time than the small ones. For companies that aim to compete with delivery speed like Ninja Van, they would prioritize the use of small vehicles (often many) to reduce work span. Therefore, we believe the convergence condition for the proposed subgradient algorithm is not very restrictive for our application.

3 Implementation and Field Study

We convinced the management team of Ninja Van with promising simulation results to conduct a field experiment in a major city of Southeast Asia, where a river separates the test area. 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 *control* area and stations that 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, while the control area had 40 stations and 850 drivers. Comparing what happened before and after the implementation to both the treatment and control area provided us with rigorous estimates of the benefits of the proposed zoning system (as well as the statistical significance of the results). We observe average reductions of 6.6% and 3.5% in the work span of the stations and the delivery time of the drivers, respectively. Based on these results and a back-of-the-envelope calculation, we estimate a total savings of 16,305,960 Thai Bhat (or equivalently, USD$440,260.92) annually in operating costs across the country of the focal market. The reduction in delivery lead time naturally increases customer satisfaction as a result of the higher level of delivery service quality. Overall, our proposed zoning model is effective in reducing delivery lead time and helps the company better distribute workload among its drivers and limit long working hours, creating a win-win scenario for both the company and its drivers.

References

Beardwood, Jillian, Halton, John H, & Hammersley, John Michael. 1959. The shortest path through many points. *Pages 299–327 of: Mathematical proceedings of the Cambridge philosophical society*, vol. 55. Cambridge University Press.

Carlsson, John Gunnar, Carlsson, Erik, & Devulapalli, Raghuveer. 2016. Shadow prices in territory division. *Networks and Spatial Economics*, **16**, 893–931.

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

Pavone, Marco, Arsie, Alessandro, Frazzoli, Emilio, & Bullo, Francesco. 2011. Distributed algorithms for environment partitioning in mobile robotic networks. *IEEE Transactions on Automatic Control*, **56**(8), 1834–1848.

Sykes, Pam. 2023. *Last mile delivery costs — Challenges and solutions.* Available at `https://www.routific.com/blog/last-mile-delivery-costs`. (Accessed December 13, 2023).

United Press International. 2020. *Delivery drivers working to death amid online shopping boom in S. Korea.* Available at `https://www.upi.com/Top_News/World-News/2020/12/24/Delivery-drivers-working-to-death-amid-online-shopping-boom-in-S-Korea/1141608844270/`. (Accessed December 14, 2023).

Zhong, Hongsheng, Hall, Randolph W, & Dessouky, Maged. 2007. Territory planning and vehicle dispatching with driver learning. *Transportation Science*, **41**(1), 74–89.