



# Urban Last Mile Delivery Data Mining for Performance Improvement

Nigel W.Z. Soh

Singapore Institute of Manufacturing Technology,  
Singapore  
nigel.soh@u.nus.edu

Fangyuan Wan

Singapore Institute of Manufacturing Technology,  
Singapore  
wan\_fangyuan@simtech.a-star.edu.sg

Allan N. Zhang

Singapore Institute of Manufacturing Technology,  
Singapore  
nzhang@simtech.a-star.edu.sg

Chi Xu

Singapore Institute of Manufacturing Technology,  
Singapore  
cxu@simtech.a-star.edu.sg

## ABSTRACT

This paper presents an in-depth exploration of data mining techniques aimed at optimizing the operational efficiency of urban last mile delivery services. Leveraging an authentic industry dataset graciously provided by a collaborative logistics partner, this study meticulously unravels intricate delivery patterns and discerns clusters attributed to delays, employing advanced cluster analysis methodologies through the utilization of the WEKA software suite. From our initial cluster analysis, we identified specifically that 33% of late cases occurred in the latitude range of (1.277, 1.287], 36% occurred in the longitude range of (103.843, 103.855], and that Driver 2065 was involved in 13% of late cases. Furthermore, a pioneering route analysis paradigm is introduced, elucidating an implementation framework harnessed through Python, Pandas, Folium packages, and the Open Source Routing Machine (OSRM) API. Through our route analysis, we were able to visualize the historical routes taken by drivers and the recommended routes by OSRM for their given jobs. In the case of Driver 2065, this allowed us to identify visits to non-job locations and extended durations spent at high-rise and high-density buildings. Notably, this research surmounts the challenge posed by imprecise GPS coordinates for job locations by propounding an innovative approach to location estimation. This groundbreaking technique bestows the capability to compute pivotal parameters, encompassing travel time and service duration, which aptly characterizes the temporal allocation at each discrete job locale. The culmination of our scholarly pursuits begets profound insights, effectively serving as a guiding compass to engender tangible operational enhancements and methodical finesse in the domain of delivery operations, thereby ensuring the punctilious execution of time-sensitive deliveries.

## CCS CONCEPTS

• Applied computing; • Operations research; • Transportation;



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICCMB 2024, January 12–14, 2024, Singapore, Singapore

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1665-2/24/01

<https://doi.org/10.1145/3647782.3647796>

## KEYWORDS

Urban Last Mile Delivery Route Analysis, Cluster Analysis, Data Mining, Decision Support

### ACM Reference Format:

Nigel W.Z. Soh, Allan N. Zhang, Fangyuan Wan, and Chi Xu. 2024. Urban Last Mile Delivery Data Mining for Performance Improvement. In *2024 7th International Conference on Computers in Management and Business (ICCMB 2024)*, January 12–14, 2024, Singapore, Singapore. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3647782.3647796>

## 1 INTRODUCTION

The field of Data Mining has emerged as a vital tool for deriving meaningful insights from vast datasets, facilitating informed decision-making across a spectrum of enterprise and supply chain scenarios. For instance, its application extends to optimizing inventory management, fuel consumption, analysing oil price movements, and the prediction of oil production, among others [1] [2] [3] [4]. Particularly pertinent to delivery-oriented businesses, data mining plays a pivotal role in enhancing job performance, yielding benefits such as heightened customer satisfaction and curbed economic losses arising from job failures. These enhancements are of particular significance in the realm of Last Mile Logistics, as this phase of the supply chain often proves to be the least efficient and cost-intensive [5]. Amid escalating market competition [6], operational efficiency becomes a cornerstone for cost reduction and the preservation of customer satisfaction [7], thereby safeguarding revenue streams.

In this context, our study is driven by the imperative to unearth actionable insights that can systematically enhance delivery performance, focusing on a real-world industry case. Our dataset originates from a third-party logistics service provider (3PL) encompassing their fleet of vehicles and drivers. Our investigation centers on identifying the causes behind delayed delivery jobs, spanning both pick-up and drop-off tasks. The 3PL's diverse vehicle fleet comprises vans, cars, and motorbikes, catering to a range of goods, from letters to medium-sized parcels. Each vehicle embarks from the central depot at predetermined times, traversing designated job locations for pick-ups and drop-offs. An internally developed Vehicle Routing Engine guides the drivers, its routes adapting dynamically based on real-time job completions.

The dataset is bifurcated into two core components: job attempt information and GPS data. The former entails crucial details like

job ID, latitude, longitude, timestamp, driver ID, and completion status, while the latter encompasses driver ID, latitude, longitude, and timestamp.

Our approach embraces both cluster analysis and route analysis. For cluster identification, we harness the user-friendly WEKA software, which streamlines dataset filtering and visual distribution representations. For route analysis, we undertake a simulation employing Python, Pandas, Folium packages, and the Open Source Routing Machine (OSRM) API. A salient feature of our approach lies in our innovative route analysis methodology. Given that the GPS data may lack precise job location coordinates, we devised a method to infer these locations, thereby enabling the segmentation of GPS data into vehicle routes between successive job locations. Moreover, this allowed us to estimate travel times between job sites and service times, reflecting the duration spent at each location.

## 2 LITERATURE REVIEW

### 2.1 Cluster analysis

Cluster analysis encompasses several techniques [8] [9], notably K-means clustering [10], hierarchical clustering [11], and density-based clustering. Of particular interest in our study is density-based clustering, which defines clusters over contiguous regions with high point densities [12]. In our study, we employ a density-based approach to discern clusters within visualizations of the distributions of datapoints from our dataset, which are generated using the WEKA software.

### 2.2 Route analysis

To gain clearer insights into vehicle travel patterns and facilitate more in-depth analysis, it becomes essential to pinpoint job locations within the GPS data. Notably, Ma, Xiaolei, et al. [13] introduced anchor points for truck trip chains through a spatial, density-based clustering algorithm on GPS data, successfully identifying frequently visited points by their fleet. However, our context differs significantly, as we focus on the entire paths traversed by individual vehicles following unique travel plans. In contrast, Ma et al.'s study centered on common points frequented by their fleet, rendering their method less applicable to our scenario. Another relevant study by Sharman, Bryce W., and Matthew J. Roorda [14] explored hierarchical agglomeration and partitioning clustering methods to ascertain trip destinations, akin to our context of individual vehicle visits to planned destinations. Yet, disparities emerge concerning dataset characteristics. While they inferred trip ends through vehicle engine on-off status or vehicle stationary periods, our dataset lacks engine status and vehicle GPS information. Instead, our data entails driver GPS information post-vehicle parking, as drivers facilitate deliveries. This intricacy compelled us to develop an innovative approach to estimating job locations within the constrained GPS dataset.

Consequently, we devised a novel method to estimate job locations, effectively navigating the limited GPS information by leveraging job attempt data. This innovative approach ensures consistency in identifying job locations across both segments of the dataset.

## 3 DESCRIPTION OF DATASET

The dataset comprises two distinct components: 1) JOB\_INFO and 2) GPS\_DATA. The former, JOB\_INFO, encompasses job attempt details, while the latter, GPS\_DATA, comprises GPS information collected from drivers at regular intervals. Tables 1 and 2 delineate the attributes characterizing each segment, accompanied by an exemplar entry.

## 4 PROPOSED METHOD

We use CRISP-DM methodology for this study. The details are described as the following subsections.

### 4.1 Preprocessing

The initial dataset, JOB\_INFO, underwent augmentation with the inclusion of the subsequent enhancements:

- ATTEMPTED\_DAY: Intended for exploratory data analysis purposes.
- ATTEMPTED\_TIME: Incorporated for exploratory data analysis endeavors.
- IS\_LATE: Introduced to distinguish between late and non-late jobs.

The new attributes 'ATTEMPTED\_DAY' and 'ATTEMPTED\_TIME' are conveniently derived from the existing 'ATTEMPTED\_DATETIME' field. Meanwhile, 'IS\_LATE' is established based on the contents of the 'COMPLETION\_DESCRIPTION' attribute, assuming a Boolean value (True or False) contingent upon whether the term 'late' is present in the corresponding entry. In addition, entries featuring 'LAT' = 0 or 'LONG' = 0 were removed, deeming them as negligible anomalies. The configuration of the modified JOB\_INFO, along with an illustrative entry, is delineated.

### 4.2 Cluster analysis

Subsequently, we initiated a cluster analysis using the WEKA software to unveil visualizations illuminating the distribution of features within JOB\_INFO. To distinguish between late and non-late jobs, we designated 'IS\_LATE' as the class variable. With an aim to mitigate the occurrence of late cases, we delved deeper by focusing on late instances. Our exploration then shifted towards identifying features with conspicuous spikes in their visual patterns, indicating potential clusters of late jobs under specific conditions. Notably, we directed our attention towards 'LAT', 'LONG', and 'DRIVER\_IDENTIFIER' features, probing the interplay among late jobs, individual drivers, and geographical coordinates.

### 4.3 Route analysis

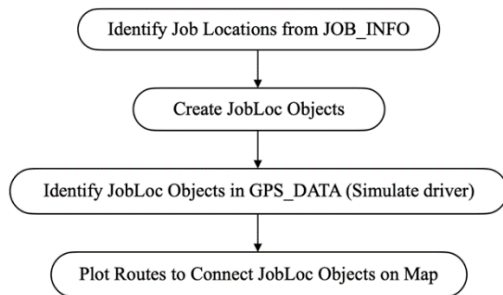
In the preceding section, our focus successfully narrowed down instances of late jobs to specific drivers and geographical regions, as defined by latitude and longitude ranges. To glean deeper insights from the available dataset, we embarked on a route analysis specifically targeting these drivers and regions. This endeavor involved examining the routes traversed by these drivers between job locations within the regions prone to late deliveries and their associated circumstances. The overall process and steps undertaken in our route analysis are outlined in Figure 1.

**Table 1: JOB\_INFO (Original)**

Feature	Example Value
JOBIDX	7338195
JOBNO	2301S00908
DELIVERYCONTENT	Parcel
ATTEMPTED DATETIME	3/1/2023 4:03:14 PM
STATUSCODEIDX	49
FROMDATETIME	3/1/2023 3:00:00 PM
TODATETIME	3/1/2023 3:30:00 PM
EXTFROMDATETIME	NULL
EXTTODATETIME	NULL
COMPLETION_DESCRIPTION	Attend Late, Std Run, PM, Late <=45min
JOB_DESCRIPTION	
JOB_DESCRIPTION2	Auto SS FP (Std Run)
DRIVER_IDENTIFIER	2065
LAT	1.2852529
LONG	103.851096
ACCURACY	4250
VEHICLETYPE	Van

**Table 2: GPS\_DATA**

Feature	Example Value
ID1	599908662
DRIVER_IDENTIFIER	421
LAT	1.2814896
LONG	103.820354
DATETIME	6/1/2023 6:33:56 AM
ACCURACY	2053

**Figure 1: Steps in route analysis**

To execute the route analysis, we designed a driver’s travel pattern simulation for a single day. This procedure commenced with the identification of the driver’s job locations. Our approach involved filtering the GPS\_DATA to extract entries pertinent to the designated driver and day. Subsequently, we filtered the JOB\_INFO data by the same driver and day, thereafter sorting all relevant job attempts based on the ‘ATTEMPTED DATETIME’ feature in ascending order. We created custom ‘JobLoc’ objects to represent

individual job locations. An illustration of a ‘JobLoc’ object is presented in Figure 2, and the key attributes of this object are detailed in Table 3. Through iteration, we populated each ‘JobLoc’ object with its respective job attempts, establishing a sequence of these objects, each representing a distinct job location.

With job locations identified, our attention shifted to pinpointing those within the regions where frequent late jobs occurred, a region referred to as the ‘late rectangle’ for simplicity. Having established the relevant job locations, we proceeded to simulate the driver’s travel pattern by iteratively analyzing filtered GPS\_DATA. Given the chronological sorting of GPS\_DATA entries, an additional sorting step was unnecessary.

The determination of the driver’s arrival at a job location relied on the Haversine formula, calculating the distance between the driver’s present location (extracted from GPS\_DATA) and the job location (from the corresponding ‘JobLoc’ object). When the calculated distance fell within a predefined threshold, the driver was considered to have reached the job location. Consequently, attributes such as ‘recorded\_entry\_time,’ ‘recorded\_entry\_loc,’ and ‘recorded\_entry\_idx’ were assigned to the respective ‘JobLoc’ object based on the current GPS\_DATA entry. Subsequent iterations marked the driver’s departure from the job location when the distance threshold was met again, resulting in the assignment of ‘recorded\_exit\_time,’

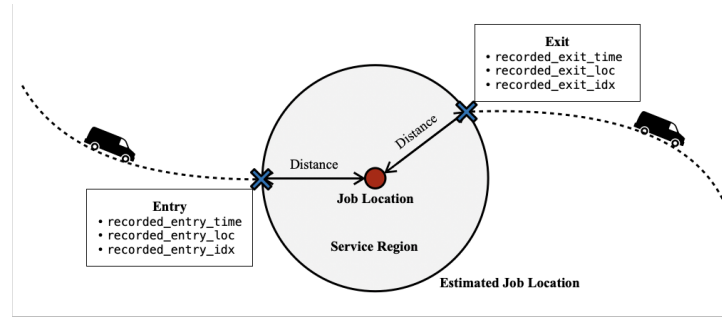


Figure 2: Illustration of estimation of a job location with a JobLoc object.

Table 3: Key Attributes of JobLoc object

Attribute	Type	Description
visit_idx	int	Contains index of JobLoc object to identify job locations and their order of completion.
lat_long	list	Stores in a list the latitude and longitude values (as floats) of a job location.
jobs	list	Stores in a list the jobs attempted at a job location.
recorded_entry_time	str	Contains the time component ('DATETIME') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver enters the job location.
recorded_exit_time	str	Contains the time component ('DATETIME') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver exits the job location.
recorded_entry_loc	list	Stores in a list the latitude and longitude components ('LAT', 'LONG') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver enters the job location.
recorded_exit_loc	list	Stores in a list the latitude and longitude components ('LAT', 'LONG') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver exits the job location.
recorded_entry_idx	int	Contains index ('ID1') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver enters the job location.
recorded_exit_idx	int	Contains index ('ID1') of the row in the GPS_DATA dataset that we estimate to be the point at which the driver exits the job location.

'recorded\_exit\_loc,' and 'recorded\_exit\_idx' attributes. By maintaining synchronization between JOB\_INFO and GPS\_DATA through the 'visit\_idx' attribute, we accurately simulated the driver's travel pattern between job locations.

The interval between entry and exit times represented the estimated service time for each job, with the circular region around the job location demarcating the service region. This cumulative information allowed us to approximate job locations within GPS\_DATA, with each 'JobLoc' object documenting these estimations. The resulting job locations were visually identified on a map, and path connections between them were plotted using the Python Folium package.

For generating a suggested route by OSRM (Open Source Routing Machine) encompassing a driver's job locations, the OSRM API was employed in conjunction with the 'lat\_long' attribute of the 'JobLoc' objects. This process entailed retrieving routes between consecutive job locations and linking them to outline the recommended path

from the initial job location within the 'late rectangle' to subsequent ones, preserving the order specified by the 'visit\_idx' attribute.

To extract the actual path taken by the driver during the day, individual paths between specific job locations of interest were determined and aggregated. The connection between two job locations was established by identifying the corresponding rows in GPS\_DATA. Each connection was visualized as a line joining the associated coordinates. To enhance visibility of driver positions along the path, markers were virtually placed at regular time intervals. By employing a predetermined interval and computing the time difference between successive GPS\_DATA rows, markers were introduced to the plot at intervals roughly equivalent to the specified time interval.

In summary, our comprehensive route analysis facilitated the estimation of job locations, visualization of drivers' paths, and the determination of suggested and actual travel routes, ultimately providing a dynamic view of drivers' activities during their journeys.

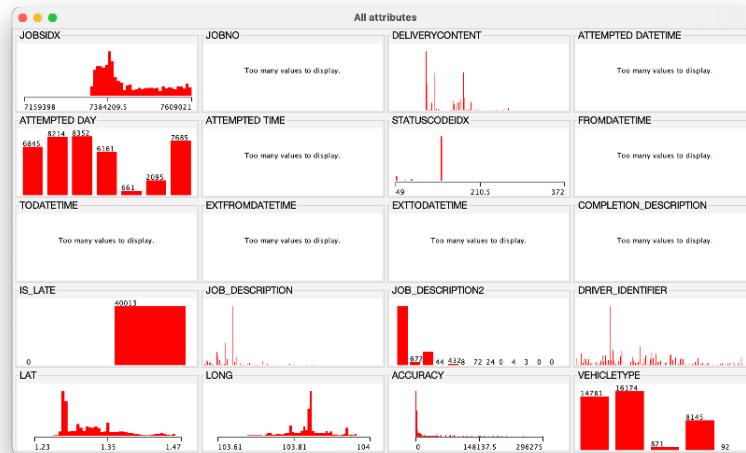


Figure 3: Filtering on IS\_LATE = True

## 5 RESULTS AND DISCUSSION

### 5.1 Results of cluster analysis

Our initial step involved filtering based on the condition 'IS\_LATE' = True, thus eliminating non-late jobs, as our focus lay in uncovering patterns within late jobs. Our intent was to delve into the interplay among late jobs, individual drivers, and geographical data, suspecting that a significant portion of late job occurrences could be attributed to locality or personnel-related issues. In this pursuit, we scrutinized the 'LAT', 'LONG', and 'DRIVER\_IDENTIFIER' features within WEKA visualizations to discern conditions giving rise to notable spikes in late cases.

It is worth acknowledging that the precise definition of a spike substantial enough to warrant our attention through visual inspection is not immediately apparent. Intuitively, such a spike might occur when a substantial proportion of cases within the overall dataset cluster within a narrow range for a specific parameter. To translate this intuition into a quantitative selection criterion, one approach involves examining the number of late cases occurring within the specified parameter range, relative to the total number of late cases. This computation is facilitated by WEKA’s visualization features, which provide case counts for specific ranges. In general, our selection of parameter ranges corresponding to significant spikes encompasses at least 10% of the total late cases. Simultaneously, comparable ranges that adhere to similar lengths would yield proportions significantly below half of this value, unless they too qualify as spikes.

As illustrated in Figure 3, substantial spikes in late cases are evident when:

1. 'LAT' lies within the range (1.277, 1.287] ( $13032/40013 = 32.569\%$  of cases)

2. 'LONG' falls within the range (103.843, 103.855] (14292/40013 = 35.718% of cases)
3. 'DRIVER\_IDENTIFIER' corresponds to 2065 (5025/40013 = 12.558% of cases)

This approach provides us with quantitatively grounded criteria for identifying notable patterns within late job occurrences, allowing us to effectively focus our investigation on specific regions and driver attributes where significant late cases cluster.

Our investigation commenced by initially filtering based on the condition 'IS\_LATE' = True, thereby excluding non-late jobs, in pursuit of discerning distinct patterns within late job occurrences. Our focus was on unveiling potential associations involving late jobs, individual drivers, and geographical attributes. As it seemed plausible that a substantial portion of late jobs might be influenced by locality or driver-related factors, we specifically examined the 'LAT', 'LONG', and 'DRIVER\_IDENTIFIER' features within WEKA visualizations. Our aim was to unravel circumstances giving rise to pronounced spikes in instances of late deliveries. It is worth noting that determining the precise magnitude of a spike meriting attention through visual inspection proved nontrivial. We sought a quantitative foundation for our selection criterion, which we achieved by evaluating the number of late cases within a specified parameter range as a fraction of the total late cases. This calculation was facilitated by WEKA's inherent visualization capabilities, offering insights into case counts across parameter ranges.

In general, the parameter ranges delineating significant spikes encompassed at least 10% of the total late cases. Conversely, parameter ranges of similar lengths yielded proportions notably below half of this benchmark, unless they also constituted spikes.

This analysis indicated that certain conditions lead to higher likelihoods of late deliveries:



- Jobs occurring within latitude range (1.277, 1.287] or longitude range (103.843, 103.855] demonstrate increased lateness.
- Jobs attributed to Driver 2065 are more prone to being late.

In a subsequent refinement, we focused on the latitude range (1.277, 1.287] as guided by the prior insight. The analysis unveiled that the most substantial spikes in late cases occurred within specific sub-ranges:

- 'LAT' = (1.278, 1.28] (47.869% of cases)
- 'LAT' = (1.285, 1.286] (21.100% of cases)
- 'LONG' = (103.849, 103.851] (50.540% of cases)
- 'DRIVER\_IDENTIFIER' = 2065 (31.844% of cases)

Further scrutiny within the 'LAT' = (1.278, 1.28] range revealed additional insights:

- 'LONG' = (103.845, 103.853] (70.249% of cases)
- 'DRIVER\_IDENTIFIER' = 2065 (32.258% of cases)

Similarly, a detailed assessment within the 'LAT' = (1.285, 1.286] range led to the following revelations:

- 'LONG' = (103.849, 103.852] (93.435% of cases)
- 'DRIVER\_IDENTIFIER' = 2065, 2169, 2811 (74.407% of cases)

Additionally, focusing on the longitude range (103.843, 103.855] unveiled further insights:

- 'LONG' = (103.848, 103.849] (26.152% of cases)
- 'LONG' = (103.85, 103.851] (44.199% of cases)
- 'DRIVER\_IDENTIFIER' = 2065 (29.961% of cases)

In summary, among late jobs:

- A significant number of jobs within latitude range (1.278, 1.28] and longitude range (103.848, 103.851] are carried out by Driver 2065.
- A noteworthy percentage of jobs within latitude range (1.285, 1.286] and longitude range (103.848, 103.851] are executed by Driver 2065, Driver 2169, and Driver 2811.

This analysis showcases a strategy for identifying prominent late regions among drivers in the dataset. A similar approach can be iteratively employed to unveil secondary late regions by consecutively eliminating late orders from higher-frequency late regions, and subsequently subjecting the remaining jobs to clustering analysis. This iterative process aims to capture all significant late regions. By leveraging this methodology, we can conduct route analysis on jobs within these regions, further enhancing the precision and effectiveness of our analysis.

## 5.2 Results of route analysis

Presenting the plotted trajectories suggested by OSRM and the actual paths traced from GPS\_DATA for Driver 2065, Driver 2169, and Driver 2811 on January 20, 2023, we offer the visual representations depicted in Fig. 4, 5, 6, and 7.

In addition, our analytical capabilities extend to evaluating travel time through insights derived from the OSRM API's recommendations. As for the actual path, a straightforward estimate of travel time can be readily attained by computing the time difference between the 'recorded\_entry\_time' and 'recorded\_exit\_time' attributes associated with the corresponding JobLoc objects. Upon meticulous scrutiny of Figure 4 and Figure 5 pertaining to Driver 2065, a conspicuous discrepancy surfaces within the designated late

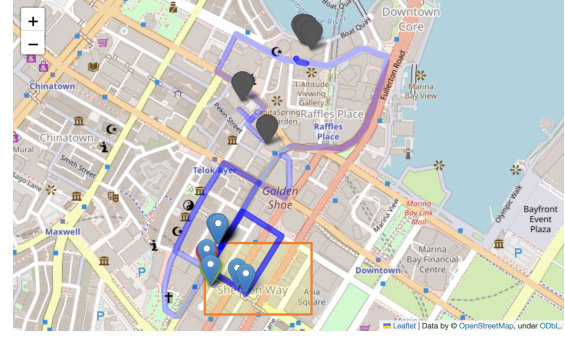


Figure 4: Driver 2065 - suggested path

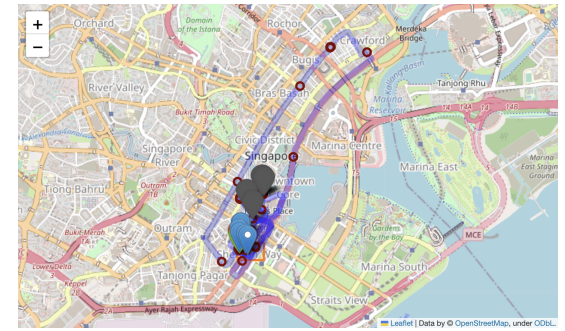


Figure 5: Driver 2065 – actual path

rectangle, encompassing the latitude range of (1.285, 1.286] and the longitude range of (103.848, 103.851]. Notably, Driver 2065's actual trajectory markedly diverges from the suggested OSRM path. This divergence primarily stems from Driver 2065's deliberate deviations, leading to a detour to non-job locations, notably conspicuous within the Crawford area positioned between job sites within the late rectangle.

A closer examination of the map reveals prominent clusters of red circles, clearly demarcating instances where the driver spent significant amounts of time at locations. These clusters notably manifest around prominent CBD landmarks, such as the Wing On Life Building and the UOB Plaza towers, prominently illustrated in Figure 6 and Figure 7 respectively. This analytical insight penetrates the intricate dynamics governing the driver's movements, casting light on recurrent deviations from the prescribed path, characterized by:

- Visits to non-job locations
  - Extended durations spent at high-rise and high-density buildings.

Conversely, a comparable finding emerged for Driver 2169, where prolonged stays at high-rise and high-density buildings feature as a key observation.

Significantly, a parallel observation is unveiled for Driver 2811, reinforcing the recurrent pattern of extended durations spent at such high-rise and high-density locations. These key findings furnish invaluable insights into the drivers' behaviors, revealing their

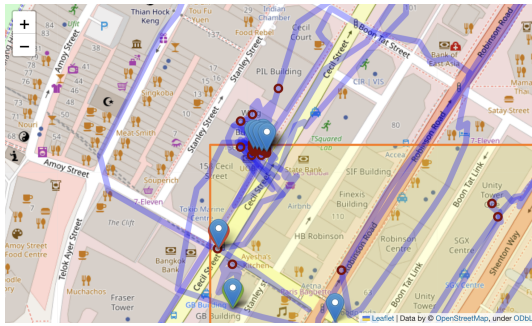


Figure 6: Driver 2065 - cluster at Wing On Life Building

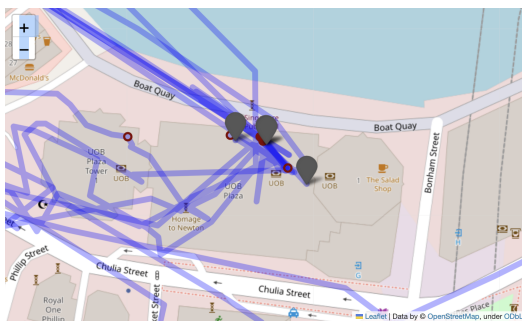


Figure 7: Driver 2065 - cluster at UOB Plaza towers

propensity to deviate from optimal routes and linger around specific areas. Such insights hold the potential for optimizing delivery efficiency and refining operational strategies to enhance the overall service quality.

### 5.3 Discussion

A common theme for the drivers was the visiting of high-rise, high-density job locations. The drivers typically spent much time at these places, judging by the density of red circles in the plots. Also it could be the case that vertical travel time (i.e. time for drivers travelling up and down the buildings) could be very long in these locations. This long and uncertain vertical travel time for intensive jobs may be a major cause for high frequency of late jobs there.

The actual total travel time of the drivers from their first job to their last job in the late rectangle is significantly greater than that suggested by OSRM. This suggests that much time was spent travelling by the drivers outside of travel times between job locations. This supports the notion that drivers might be spending much time travelling within buildings – something that OSRM cannot capture. Another common finding was that the drivers often visited non-job locations between jobs in the late rectangle. Naturally, this might cause delays subsequently and late jobs. Interestingly, many of these non-job locations are food establishments. The drivers tend to visit such places in the early afternoon, suggesting that they might be visiting these places to buy their lunch.

## 6 CONCLUSION AND FUTURE WORK

This paper presents an innovative approach to estimating job locations from GPS data within a comprehensive data mining framework aimed at enhancing delivery performance for a logistics partner. Notably, the analysis highlights that late jobs are influenced by factors such as high-rise, high-density locations and visits to non-job sites. Addressing these factors holds the potential to reduce late deliveries and enhance overall performance for the logistics partner.

Future work could refine our approach by developing an adaptive distance threshold to address the accuracy impact of overlapping service regions in our method. Mitigating this challenge requires careful threshold calibration for optimal simulation accuracy without underestimating service times. Additional opportunities include creating statistical models for location-dependent service times and non-job events (e.g., lunch breaks) based on historical data. Integrating these models into logistics planning can enhance delivery operation efficiency.

## ACKNOWLEDGMENTS

This work was in part supported by the IEO Decentralised GAP project of Continuous Last-Mile Logistics (CLML) at SIMTech under Grant I22D1AG003. The authors thank XDel Singapore for providing the data and feedback.

## REFERENCES

- [1] Pujiarto, Bambang, *et al.* "A data mining practical approach to inventory management and logistics optimization." *International Journal of Informatics and Information Systems* 4.2 (2021): 112-122.
- [2] Muchová, Miroslava, Ján Paralič, and Michael Nemčík. "Using predictive data mining models for data analysis in a logistics company." *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology-ISAT 2017: Part I*. Springer International Publishing, 2018.
- [3] Kwan-Hua Sim, Nicholas Ching-Yun Bong, and Kwan-Yong Sim, "Analyzing Price Movement of Crude Oil through Historical Price Data Distribution by Using Apriori Data Mining Algorithm," *International Journal of Computer Theory and Engineering* vol. 9, no. 5, pp. 334- 338, 2017.
- [4] Zahra Mahdavi, Maryam Khademi, "Prediction of Oil Production with: Data Mining, Neuro- Fuzzy and Linear Regression," *International Journal of Computer Theory and Engineering* vol. 4, no. 3, pp. 446-447, 2012.
- [5] Gevaers, Roel, Eddy Van de Voorde, and Thierry Vanelslander. "Characteristics and typology of last-mile logistics from an innovation perspective in an urban context." *City distribution and urban freight transport: Multiple perspectives* January (2011): 56-71.
- [6] Lopez, Edwin. "Last-Mile Delivery Options Grow Ever More Popular." *Supply Chain Dive*, 11 May 2017, [www.supplychaindive.com/news/last-mile-technology-options-investment/442453/](http://www.supplychaindive.com/news/last-mile-technology-options-investment/442453/).
- [7] Burity, Jasminea. "The importance of logistics efficiency on customer satisfaction." *Journal of Marketing Development and Competitiveness* 15.3 (2021): 26-35.
- [8] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Data mining cluster analysis: basic concepts and algorithms." *Introduction to data mining* 487 (2013): 533.
- [9] Kishana R. Kashwan and C. M. Velu, "Customer Segmentation Using Clustering and Data Mining Techniques," *International Journal of Computer Theory and Engineering* vol. 5, no. 6, pp. 856-861, 2013.
- [10] Kriti Srivastava, R. Shah, D. Valia, and H. Swaminarayan, "Data Mining Using Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment," *International Journal of Computer Theory and Engineering* vol. 5, no. 3, pp. 520-522, 2013.
- [11] Dongkuan Xu and Yingjie Tian. 2015. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* 2, 2 (2015), 165-193. DOI:<https://doi.org/10.1007/s40745-015-0040-1>
- [12] Claude Sammut and Geoffrey I. Webb. 2011. *Encyclopedia of Machine Learning*. Springer Science & Business Media.
- [13] Ma, Xiaolei, *et al.* "Understanding freight trip-chaining behavior using a spatial data-mining approach with GPS data." *Transportation Research Record* 2596.1

(2016): 44-54.

- [14] Sharman, Bryce W., and Matthew J. Roorda. "Analysis of freight global positioning system data: Clustering approach for identifying trip destinations." *Transportation research record* 2246.1 (2011): 83-91.