Development of Large-scale Trip Analysis Toolkits for Vehicle-based GPS Trajectories using Apache Spark and Open Data: A Case Study of Taxis in Bangkok, Thailand

Apichon, W.* and Athispat, P.
School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12120, Thailand
E-mail: apichon@siit.tu.ac.th,* m6722040083@g.siit.tu.ac.th
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v21i1.3783

Abstract
Urban planning and mobility analysis have traditionally been studied through observation or questionnaires, which can be time consuming and costly. However, the rapid advancement of technology has enabled tracking devices to be installed in individual vehicles, allowing the measurement of various values, particularly global positioning system (GPS) signals. The location data collected is accurate, regularly updated, and can offer valuable insights into people's movements and behavior. Because the amount of trajectory data is substantial and continues to increase over time, specialized platforms and skills are needed for its analysis. In this study, we developed large-scale analysis toolkits to extract insights, including trip statistics, origin–destination analysis, and hotspot identification from vehicle-based GPS trajectories. The toolkits are specifically designed to handle large-scale datasets using Apache Spark, an analytics engine capable of processing large volumes of data by distributing tasks across a Hadoop cluster for efficient processing. Algorithms for the analytics model were created to reconstruct trips based on their type of mobility, and trip locations were mapped using open data such as administrative boundaries and points of interest. We then verified our approach using real-world taxi data from Bangkok, Thailand. The results revealed that taxis had more vacant trips than busy trips, and the travel time and distance taken to search for passengers was longer than those taken to pick them up and drop them off. Taxi activity was concentrated in the city center and nearby areas, particularly those within the vicinity of transport-connecting hubs. Taxi stay hotspots were mainly areas near tourist attractions and parking hubs. Furthermore, we found that the processing performance of the proposed approach increased with the number of executor cores. This study comprehensively presented information on taxi travel patterns, service availability, hotspots, and processing performance using the developed trip analysis toolkits.

Keywords: Apache Spark, Big Data, Mobility Analysis, Open Data, Taxi Probe Data

1. Introduction

Urban planning and infrastructure management are traditionally studied using data from observations or questionnaires, which are relatively time consuming and costly. Data on public transportation in a specific region necessitates collection from many sources, such as drivers, users, and stakeholders, which takes some time and have relatively small sample sizes [1]. Questionnaires can be excessively subjective because respondents may not pay sufficient attention during the response process. With the rapid development of technology, tracking devices have been installed on individual vehicles to measure various values, especially global positioning system (GPS) signals. Information obtained from these devices is straightforward, accurate, and readily available. Therefore, when information is updated, devices can manage data changes more appropriately and rapidly than traditional methods.

Mobility analysis is a method for understanding and modeling people's mobility in terms of when, where, and how they move from one place to another [1]. Trajectory data obtained from tracking devices can be used in mobility analyses to generate indicators for various aspects, including economic, environmental, and social factors. The economic dimension can be used to describe the cost of transport services such as parking fees or tolls. The quality of transportation for disadvantaged individuals can be considered from a social perspective.

In the environmental dimension, energy consumption by mode of transport can be used to analyze the pollution released by vehicles [2]. Recently, many public transport vehicles, such as buses, trains, ships, air, and taxis, have become equipped with GPS tracking devices for security reasons. Their probe data have been collected over a long period of time and can be used for mobility analysis. In Bangkok, few studies have been conducted on probe data analysis. A study utilized a questionnaire survey and probed taxi data to analyze the characteristics of taxis and service behavior [3]. Statistical analyses of taxi trips, considering trip volume, distance, duration, revenue gain rate, and origin destination (OD), have also been performed [4] and [5].

Multiple perspectives must be considered when analyzing probe data. The first is the algorithm, which enables the conversion of raw GPS data into more meaningful and easy-to-use data. The second is the platform for handling large-scale datasets, which continues to expand. The third is the set of useful indicators, which can be derived from the dataset, and the last is the software, which allows developers to develop, customize, and run the entire dataset efficiently. Many researchers have focused on specific perspectives without considering the entire flow of processing and analysis. For instance, analyses mainly focused on the moving period by using the "for hire" light flag to indicate taxis' occupancy or vacancy status, which cannot be applied to the dataset without this feature [4] and [5]. However, inaccuracies in "for hire" light data may arise due to the non-activation of meters in some taxis. According to the complaint handling statistical report for fiscal year 2024 from the Department of Land Transport, 1,797 fare meters in taxis were not used [6]. Moreover, other aspects aside from the moving period should also be considered, such as stop periods and travel behaviors, which were not mentioned in previous studies. In addition, these studies did not investigate the use of a large-scale data-processing platform. Analyzing the trajectory data requires techniques or algorithms that generally use GPS data to create trips. One of the challenges in trip construction is distinguishing stop points or areas where time is spent because GPS coordinates are expressed in the form of latitude and longitude, which change over time and are difficult to understand. Therefore, identifying stop points is the first step in transforming raw GPS data into smaller trip data that are easier to understand [7]. The generated trips can be further utilized to create mobility indicators. During the moving periods, the general indicators were the daily volume, distance, duration, and OD points. As datasets are relatively large and expand rapidly, a large-scale platform that can help process probe data is necessary to produce actionable insights. One such platform is Apache Hadoop [8], which supports the scalable storage and distributed processing of multiple nodes in a cluster. Apache Spark [9] is an open-source distributed data-processing framework designed for large-scale applications that provides a unified engine for batch and real-time data processing, which can be used for data engineering, data science, and machine learning on single-node machines or clusters. Spark performs in-memory computations, making data processing faster than the traditional Hadoop MapReduce approach, which reads or writes data on a disk [10]. Apache Spark with Pyspark can be used to improve processing performance and support easy development. However, a combination of algorithms and software developed to operate on large-scale platforms is still required.

In this study, we proposed toolkits that can be used to generate mobility indicators using probe data. The generated indicators cover the dimensions of travel patterns, accessibility, speed, safety, and stay patterns. In addition, we applied administrative boundary data to make the indicators more meaningful. The toolkits were implemented on the big-data platform, Apache Hadoop and Spark, with Pyspark, which has been widely used to build open-source toolkits and can provide better performance than traditional processing. The main contributions of this study are the 1) design and development of an open-source toolkit for generating mobility indicators, 2) construction of large-scale trips with efficient execution using Apache Spark, 3) proposal of mobility indicators that cover travel patterns, accessibility, speed, safety, and stay patterns, and 4) validation of the toolkits using real-world taxi data from Bangkok, Thailand.

The remainder of this paper is organized as follows: Section 2 describes the materials and methods. Section 3 presents the results of the analysis and the performance of Spark. Finally, Section 4 concludes the paper and discusses future research directions.

2. Data and Methods

2.1 Data Structure and Format

Toolkits were designed to process and analyze trips and mobility using the taxi GPS trajectories. The structure and format were derived from the taxi data provided by the Thai Intelligent Traffic Information Center, with the unnecessary data columns removed. The data features (vehicle ID, latitude, longitude, timestamp, and "for hire" light status) and their formats are listed in Table 1.

Table 1: Description of the probe taxi trajectory data

Feature	Description	Example
Vehicle ID	Unique vehicle ID	7LO1lqJvO7t0Scmfc...
Latitude	GPS latitude up to 5 decimal places	14.51006
Longitude	GPS longitude up to 5 decimal places	101.3771
Timestamp	GPS timestamp (yyyy-MM-dd H24:mm:ss)	2023-01-28 12:38:14
For_hire_light	"For Hire" light is ON 1 = light on => possibly no passenger 0 = light off => possibly carrying passengers	0

Table 2: Mobility indicators used in the study

No.	Group	Indicator	Definition
1	Travel Pattern	Total trips [number]	Reflection of the mobility demand in the period of interest. In general, studying within the daily range is preferred to analyze the typical daily trip volume of users.
2	Travel Pattern	Trip length (or distance) [km]	An aggregate sum of consecutive GPS points in the trip. It shows the characteristics of the passenger route on a busy trip and indicates how far a taxi usually travels to search for a customer on a vacant trip. The total distance of the trip is calculated as follows: $D_{ij} = d_{i,i+1} + d_{i+1,i+2} + d_{i+2,i+3} + \ldots + d_{j-1,j}$ where $d_{i,i+1}$ is the distance between the current point (i) and the next point (i+1), and i and j indicate the start and end points of the trip, respectively.
3	Travel Pattern	Travel time [min]	The interval between the start and end points of the trip. It indicates the duration a passenger spends on a busy trip, and the duration of a vacant trip indicates how long a taxi takes to search for a customer. Assuming that a trip has n consecutive points (P_1, P_2, P_3, …, Pn), the travel time is calculated as $T_{trip} = t_n - t_1$, where t_1 and t_n are the times at P_1 and P_n, respectively.
4	Travel Pattern	Origin–destination	The origin and destination locations of a busy trip, with the start point as the origin (pickup) and the endpoint as the destination (drop-off). Assuming that a trip consists of n consecutive points (P_1, P_2, P_3, ..., P_n), P_1 is the origin and P_n is the destination.
5	Accessibility	Service availability map (pick up / drop-off)	A density map representing the concentration of points in an area. A high-density area is a specific location that consists of a large number of points and a low-density area has a lower number of points. The boundary of each density map depends on the distribution of points. Fewer scattered points will lead to narrow boundaries. Conversely, the boundary becomes more prominent when the points have a higher dispersion.
6	Accessibility	Accessibility hotspot	A hotspot is a specific location with the highest density of data points in the density map area. In the case of the origin (pickup) point, a hotspot represents a location that has the most taxi demand, and the destination (drop-off) point represents an end-way place where most passengers usually travel.
7	Speed and Safety	Distribution of speed on vacant and busy trips	Vehicle speed represents the average taxi speed used on both vacant and busy trips. It is calculated as $V_t = D_t / T_t$, where D_t is the total distance of the trip and T_t is the interval between the time at the start and end points of the trip.
8	Speed and Safety	Speeding area	The speed of individual points that taxis use on a specific travel path or location. It is calculated as $V_{i+1} = D_{i,i+1} / T_{i,i+1}$, where $D_{i,i+1}$ is the distance from the current point (i) to the next point (i+1) and $T_{i,i+1}$ is the interval between the current point (i) and the next point (i+1).
9	Stay pattern	Stay hourly volume	The taxi volume during the stay periods are shown as the hourly volumes of taxis within a day. This number shows the number of taxis which were stationary at each hour of the day. It is used to identify which period taxi drivers typically rest or work.
10	Stay pattern	Stay hotspot	A hotspot representing the specific area for locating where taxi drivers are likely to take a car break. It can be divided into groups by stay period and further used in the characteristics analysis of each group.

The dataset file was in CSV format, with the header as indicated. Administrative boundaries were set for the administrative-level OD analysis. The boundary data were in shapefile format. The latitudes and longitudes followed the WGS84 coordinate system. The data files were either separated by date or combined with data from multiple days.

2.2 Mobility Indicators

The indicators of mobility are travel time, cost, and the variability in travel time and cost [11]; specifically, mobility is higher when the average travel times, variations in travel times, and travel costs are low. Mobility indicators can be used for several purposes. The indicators can provide trend information from which implications for transportation can be drawn or transportation policy and investment decisions can be made. These can also provide a basis for comparisons among areas and a sense of whether system performance is improving or worsening. In conjunction with intelligent transportation system strategies, indicators can be used on a real-time basis to inform travelers of the current "mobility conditions" of the transportation system such that travel decisions can be made with full knowledge of what to expect. Using our toolkits, we developed an algorithm and software to derive multiple mobility indicators in the domains of travel patterns, accessibility, speed and safety, and stay pattern. Table 2 presents the definitions and details of these indicators.

2.3 Overall Processing

The trajectory of the GPS points based on the time series $T = \{P_1, P_2, P_3, \ldots, P_n\}$ was used to generate the trip for use in the analysis following the overall process diagram in Figure 1. The process begins by identifying the stay location from the GPS log and then creating trip segments based on the stay and move segments. In the next step, the move segment was used to extract the busy and idle taxi segments. A busy taxi delivers passengers, and can be used to analyze the origin and destination. Idle indicates that the taxi is vacant and has no passengers. The speed is calculated for each moving segment, and this was used for speeding analysis. The trip segment and reconstruction results were used to produce the indicators. Apache Spark with the PySpark library running on the Hadoop cluster was utilized to allow distributed processing and improved processing time.

2.3.1 Trip reconstruction

A trip comprises three periods: stationary (stay), busy, and vacant. A stationary segment is when a taxi visits a location and remains there or moves within the specified duration and distance thresholds. The latitude and longitude of the stay point represent the centroid coordinates of the points. The busy segment is a period when the taxi is moving and possibly carrying passengers, during which the "for hire" light is off. The vacant segment is a period when the taxi moves without carrying any passenger, during which the "for hire" light is on. The method for constructing a trip comprises two main steps.

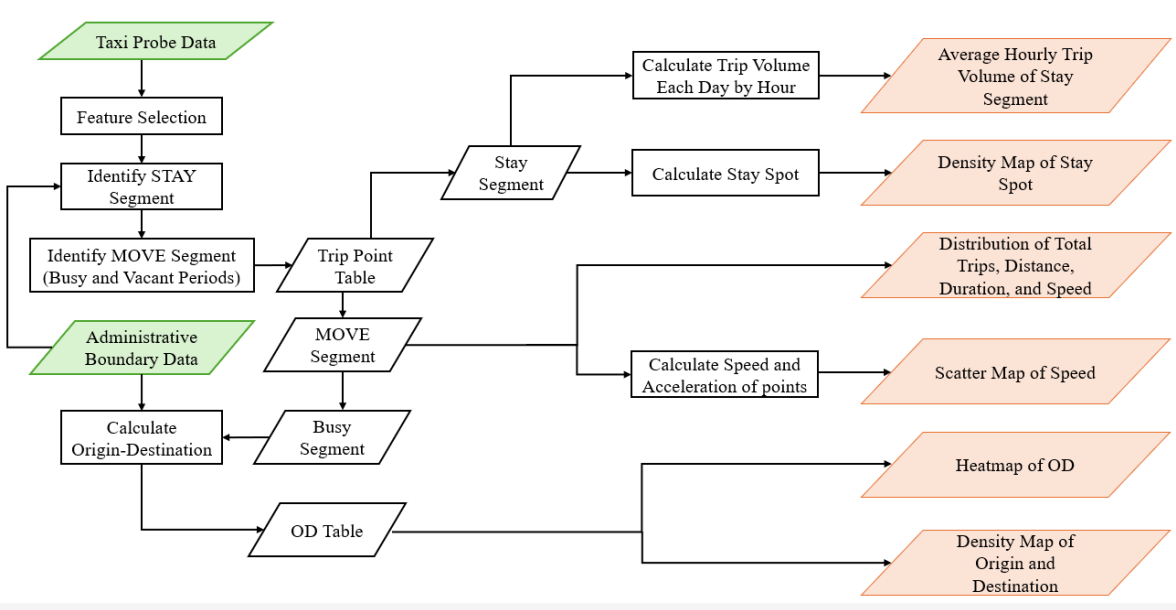

Figure 1: Overall process diagram

The first step involves identifying the stationary segment from the raw GPS data, and the second step is defining the busy or vacant periods of the move segment from among the taxi stops. Figure 2 illustrates the steps involved in trip reconstruction.

A. Identification of stop (stay) points from GPS data

Stop points can be characterized in three ways [12] geographic information, spatial clustering, and dwell-time threshold [13]. We used a stay point algorithm with two important thresholds: speed and time. Staying in one place longer than the threshold is assumed to be a stop point [14]. The appropriate threshold values depend on the characteristics of the data. This study considered the daily GPS trajectory of each taxi as a sequence for calculating the trip. The duration (δ_t) and distance (δ_d) thresholds are two thresholds we used to infer the movement status (moving or stationary) of the taxi at each point. To identify the stay points for each vehicle, we grouped the data points per day into a sequence according to the timestamp. The first point was set as the reference point ($P_{checkpoint}$) of the candidate stay point. We then checked the condition of the next points against the $P_{checkpoint}$. If the distance between the point and $P_{checkpoint}$ is shorter than the δ_d, the point is considered a member of the candidate stay point. Then, we continued to measure the distance with the point after the next point until it is over δ_d. Subsequently, we measured the overall duration starting from $P_{checkpoint}$ to the last candidate point. If the duration is greater than δ_t, it indicates the end of the current stay. Therefore, all points from $P_{checkpoint}$ to this point were used to calculate the centroid, and the information was then added to the list of stay points. Otherwise, if the duration is less than δ_t, the point is still moving; thus, we reassigned the checkpoint and repeated the method until the checkpoint was the same as the length of the trajectory. To determine the appropriate thresholds, we calculated the stay points based on different distance and duration thresholds. From Figure 3, we observed that the curves decreased intensively when δ_t increased from 0 to 500 s. When $\delta_t > 500$ s, the curves became flat. Meanwhile, for the different distance thresholds, the number of stay points was similar. Therefore, the threshold duration was set to 500 s. In addition, every distance line converges for this duration. Therefore, we chose 150 m as the distance threshold owing to the accuracy of the GPS data.

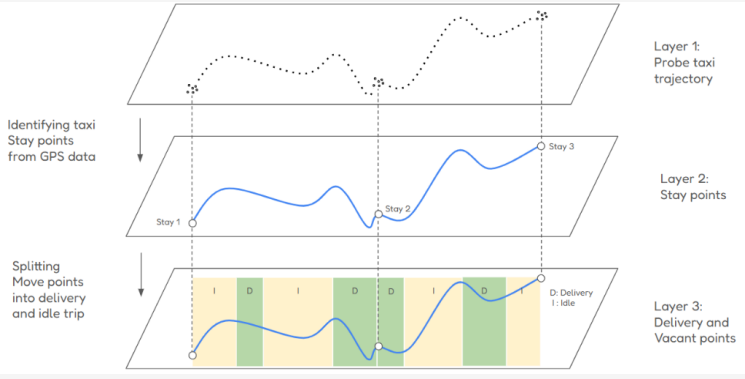

Figure 2: Overall processing from raw GPS data to trip information as three layers

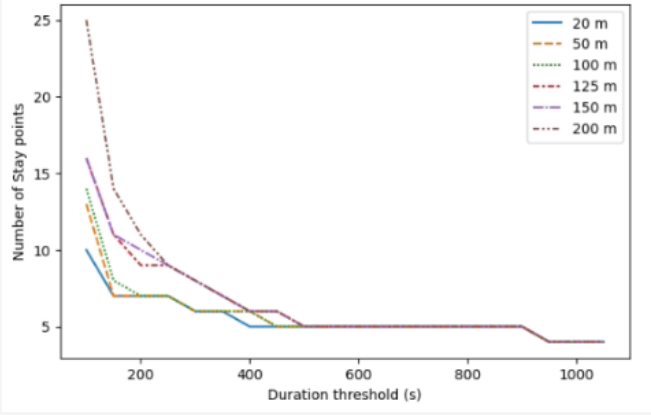

Figure 3: Number of stay points derived for the difference in distance and duration thresholds of the sample taxis

B. Identification of the busy and vacant segments

The move segment can be split into two parts based on the status of the "for hire" light feature. An active light flag suggests the lack of passengers in the vehicle, whereas passengers may be present when it is turned off. To define a busy or vacant segment, a stay point identification process was first performed. We started at the first coordinates and then moved through the trajectory; points considered within the interval of the stay period were skipped. Conversely, the remaining points were counted in the move segment. We then used the "for hire" light to classify the move type as either a busy or vacant trip.

2.3.2 OD density map

OD information can describe the mobility of a taxi carrying a passenger. After the trips were generated, busy trips were used to create the OD trip. To obtain the OD trip, we first extracted the starting point (pickup) and endpoint (drop-off) of busy trips. However, both extracted points were in longitude and latitude coordinates. Therefore, the points were mapped against Thailand's sub-national administrative boundaries to identify the regions to which they belonged. After the OD trips were identified, the origin or destination points were used to cluster the density maps. This study focused on the hourly changes in the density map by grouping points by hour. During the clustering process, the DBSCAN algorithm was utilized to eliminate noisy data because it works efficiently on noisy data. Two important arguments must be adjusted in the algorithm: Epsilon (EPS) and the minimum sample; EPS is the maximum distance between two samples for one to be considered in the neighborhood of the other. The minimum sample is the number of samples in a neighborhood for a point to be considered a core point. If it is set to a higher value, the output cluster is denser, and if it is set to a lower value, the cluster is sparser. In our study, the EPS was set to approximately 100–150 and the minimum sample size was set to approximately 25–35.

2.3.3 Characterization of stay points among the stop points

From the stay periods generated from the trip reconstruction, we illustrated the probability distribution of the stay time (Figure 4). The results suggest that the stay duration follows a broken power-law distribution and can be manually split using the breaking points at 15 and 200 min. Therefore, in this case, we assumed that the stay could be divided into three parts: <15 min, 15–200 min, and >200 min. A period >200 min likely means that the vehicle was in a long-term parking space, such as at home or at a parking station.

2.4 System Architecture

GPS data are large datasets that common computer systems and databases cannot process within an acceptable timeframe. Apache Hadoop is used to handle such large datasets with scalable features. This platform can store large amounts of data and has a high processing speed because it comprises multiple computer nodes. The actual data are split into small files and stored in different nodes.

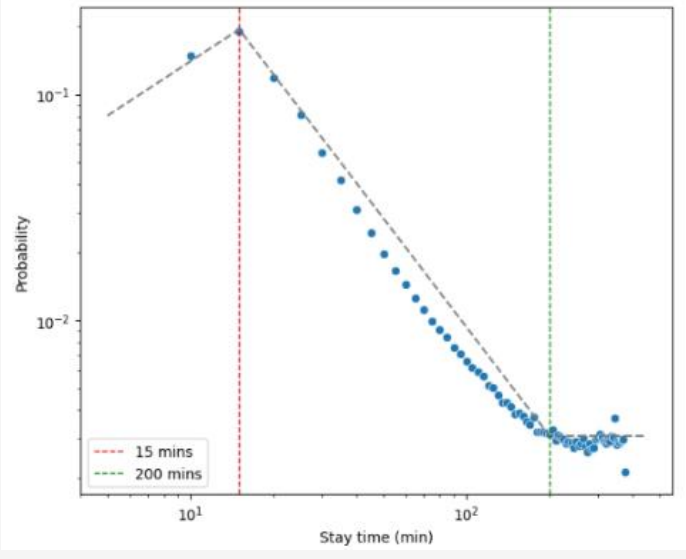

Figure 4: Distribution of taxi stop duration. The three gray dashed lines represent three fitted power-law segments. The two dashed vertical lines denote the breakpoints (15 and 200 min) in the distribution

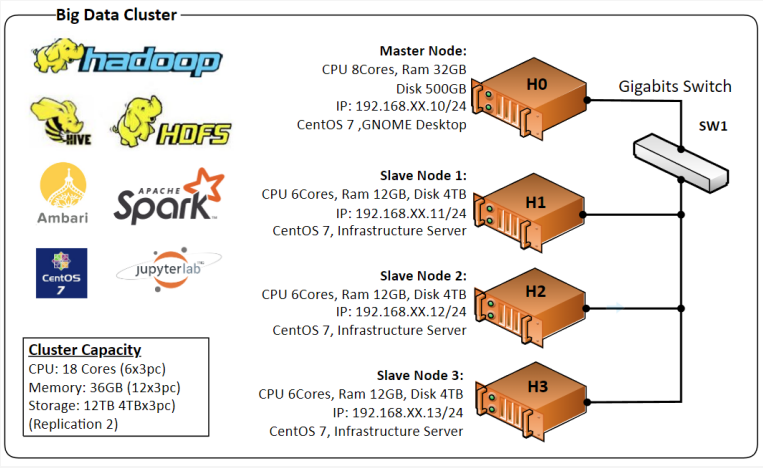

Figure 5: Hardware and software specifications of the big data cluster

Moreover, Hadoop can utilize multiple nodes for parallel processing, which increases processing speed. A big data cluster was developed based on the Apache Hadoop platform and software within its framework. Apache Hadoop was installed as the base infrastructure using Apache Ambari, a web-based management tool. Additional software, including HDFS, Hive, and Spark, were installed. Notably, Spark is a data analysis package that targets users familiar with and comfortable with Python and SQL to perform ad hoc queries, summarization, and data analysis. Additionally, it provides a mechanism for developing custom functions for specific and specification-based processing.

Figure 5 illustrates the hardware specifications and configuration of the proposed cluster. For the software part of developing the trip analysis toolkit, we utilized JupyterLab, a web-based integrated development environment with a flexible interface for users to configure and arrange workflows. The Python programming language was primarily used in the development process, requiring PySpark as the Python API for Apache Spark to perform large-scale data processing in a distributed environment. In the trip construction process, a User-Defined Function was used to create customized functions. Geopandas and Shapley were incorporated into our function to allow for geospatial processing. This software is shared on the GitHub repository and allows for public use [15]. The process involves seven steps: initialization, data importation, trip generation, OD generation, speed calculation, and visualization. Hive tables were used to store the final products of each process.

3. Results and Discussion

3.1 A Case Study of Taxis in Bangkok

To demonstrate the use of the developed toolkits, we analyzed the GPS trajectory data of taxis in Bangkok. The trajectory data were provided by the Thai Intelligent Traffic Information Center. The GPS data for taxis in Thailand were collected every three minutes for inactive cars, and every minute otherwise. The data contained nine features in the same format as our defined format (Table 1). As shown in Figure 6, the trip coordinates are distributed throughout Thailand, with a high density in the capital city, Bangkok, and other metropolitan areas. Based on the data from January 2023 (Figure 7), the average number of taxis was approximately 3000 IDs per day, and the daily volume of coordinates was approximately 1.8 million points. Notably, some days had significantly less data than others (e.g., on January 16, 17, and 19). Moreover, data for January 18, 2023 were missing. We used administrative boundaries to create the analysis results for different administrative levels. The shapefiles of the administrative boundaries were obtained from the Humanitarian Data Exchange (HDX). Shape data were classified into four levels: 0 (country), 1 (province), 2 (district), and 3 (subdistrict).

3.2 Travel Patterns

The total number of trips reflects the overall and daily mobility demands. Approximately ten trips were recorded per day. The average number of trips per taxi was the same throughout the days of the week. Even though the number of trips between vacant and busy trips were similar, an examination of Figure 8 reveals that the number of vacant trips was slightly higher than that of busy trips during four of the seven days.

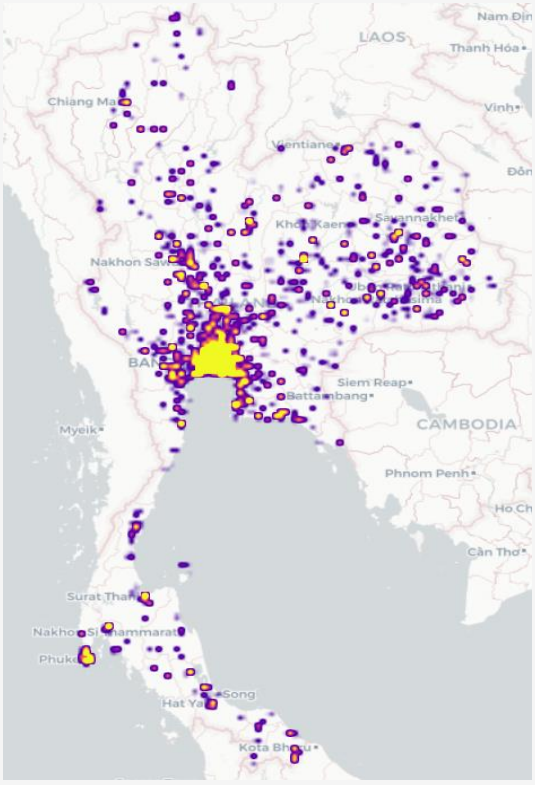

Figure 6: Distribution of Taxis within the target area

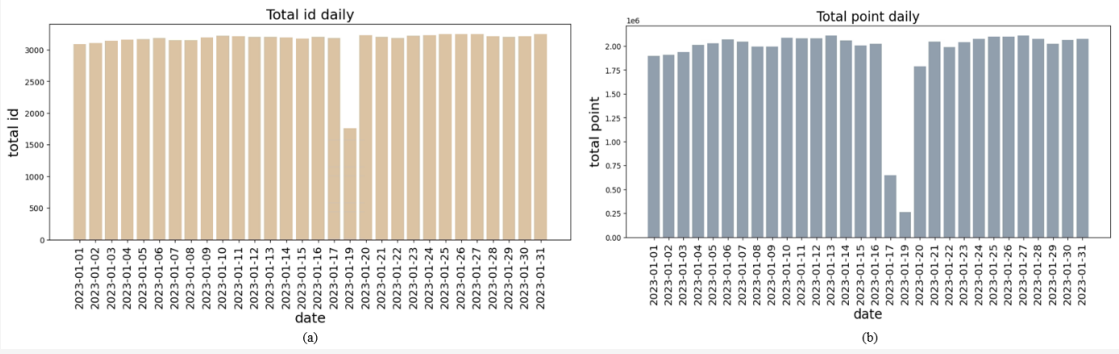

Figure 7: Statistics of the data on daily (a) total IDs and (b) total points

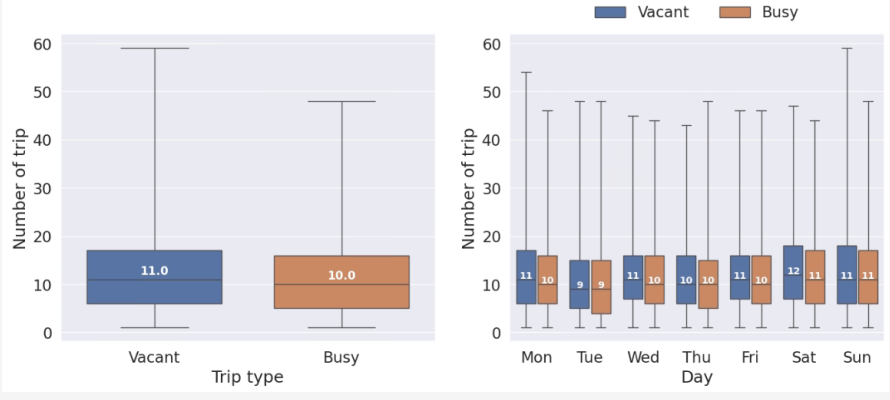

Figure 8: Average number of daily trips per taxi

The average number of vacant trips per day was approximately 11, whereas the number of busy trips was approximately 10 per day. The trip length indicates the distance a taxi usually travels to locate customers. The distributions of both vacant and busy trips were similar in shape (right-skewed), as shown in Figure 9. The median length of a vacant trip was 7.54 km, and that of a busy trip was 4.38 km. These numbers show that taxis travel farther distances to search for customers and shorter distances to transport passengers. The travel time shows the duration a passenger spends in a taxi. As shown in Figure 10, both vacant and busy trips have similar distributions and are right-skewed. The median duration of a vacant trip was 18 min and that of a busy trip was 12.75 minutes. These numbers show that a taxi takes longer to search for a customer and requires less time to transport passengers. OD points reveal the density heatmaps of origin (pickup) to destination (drop-off) regions in a specific district. A heatmap of the OD points in Bangkok is shown in Figure 11. Evidently, most of the OD movement is concentrated within the district itself or a nearby area. Chatuchak has the highest taxi demand in terms of both origin and destination because it is a transport-connecting hub for traveling in Bangkok and other provinces that supports various travel forms, such as sky trains, subways, buses, and vans.

3.3 Accessibility

Owing to the variety of origin and destination locations during each period of the day, we selected three time periods for this analysis: morning (6–7 am), afternoon (12–1 pm), and evening (5–6 pm). For the morning period (Figure 12(a)), the origin points were classified into three types: hospitals, transport hubs, and tourist attractions. Nana's tourist attractions are in a location with many hotels and attractions around the city. The transport hubs include the Suvarnabhumi Airport and Mochit 2 bus stations. Nakornchai Air Bus Station was the primary destination in the morning. For the afternoon period (Figure 12(b)), both the origin and destination points were relatively similar. The origins were mostly located around the city center, including two airports, a bus terminal, and suburban shopping malls. The destinations were centralized in the city center, such as in hospitals and shopping malls around Siam.

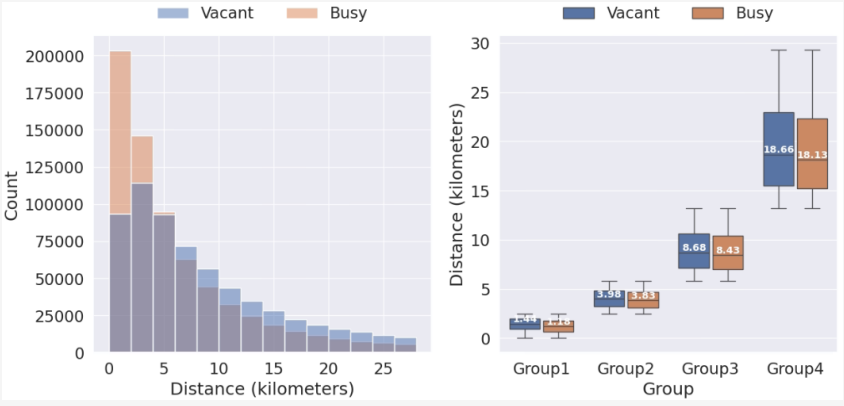

Figure 9: Distribution of trip distance

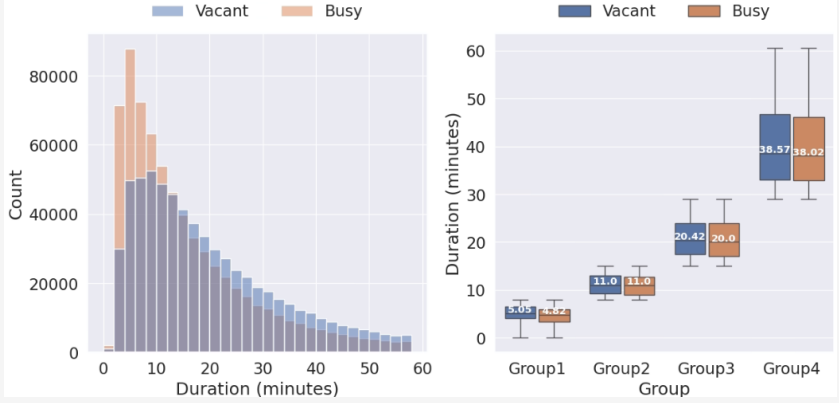

Figure 10: Distribution of trip duration

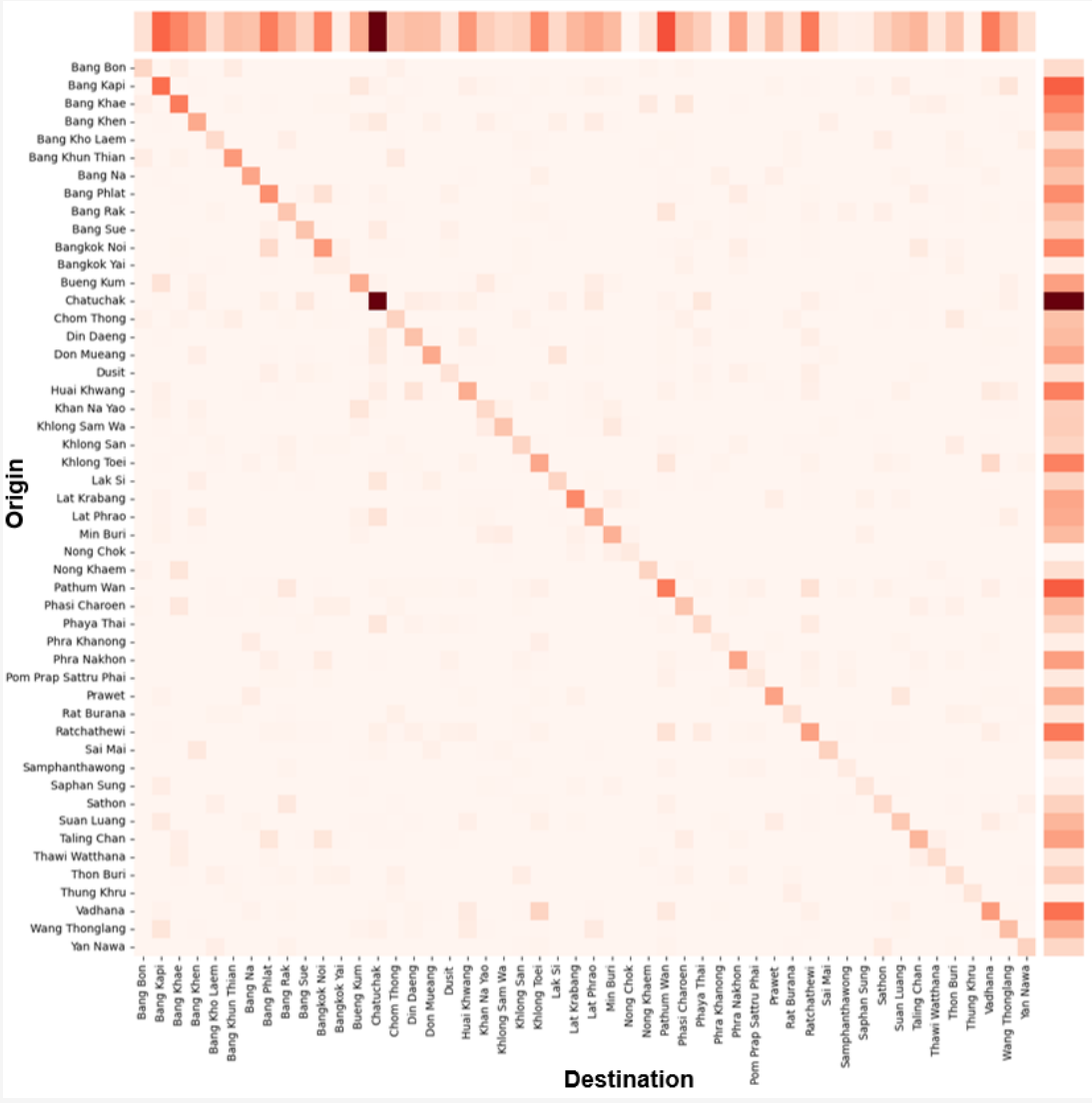

Figure 11: Heatmap of OD points in Bangkok

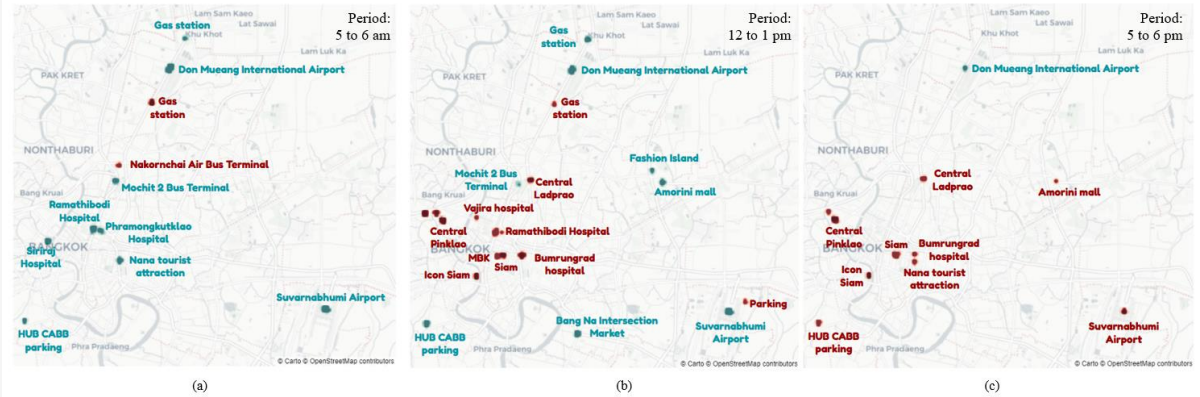

Figure 12: Density map of OD points during various periods: (a) 5–6 am, (b) 12–1 pm, and (c) 5–6 pm; origin and destination points are in green and red, respectively

In the evening (Figure 12(c)), Don Mueang International Airport was the only origin spot. In contrast, several locations were destinations, most of which were shopping malls, such as Central Pinklao, Central Ladprao, Icon Siam, and Amorini Mall. These places are transfer areas for other modes of transportation, such as Central Ladprao, which consists of two subway lines (Skytrain and subway), and Icon Siam, which consists of a Skytrain station and a pier. Other transport hubs included Siam (Skytrain station) and the Suvarnabhumi Airports.

3.4 Speed and Safety

The distribution of vehicle speed for vacant and busy trips is shown in Figure 13. Evidently, the vacant and busy trips had similar distribution. However, the median speed for vacant trips was relatively higher, with median vacant and busy speeds of 25.17 and 21.15 km/h, respectively. The speeding area displays a scatterplot of the speed range in a specific area on a map. After removing the outliers, the speeds typically used by taxis did not exceed 120 km/h.

In our example, we compared the speeding usage between a weekend (Figure 14(a)) and a weekday (Figure 14(b)) at 8 am, which concentrated only on more than the 95 percentile of speed. Taxis tended to be faster on weekends, averaging 90–120 km/h, compared with weekdays, which averaged 75–90 km/h. These data could be used to identify areas where vehicles are likely to overspeed.

3.5 Stay Patterns

As shown in the hourly trip volume of taxis in Figure 15, two different patterns of taxi behavior were observed during the weekdays and weekends for the stationary period. In the morning, taxis started earlier (approximately 7–8 am) during weekdays and later (10 am) during the weekends. On weekdays, the break periods of taxis were during the early morning (2–5 am), afternoon (12–3 pm), and late night (9–12 am). Additionally, during Fridays, more taxis tended to have no break at 6 pm compared with that during other days. On weekends, taxis usually had longer rest periods from early to late morning (2–10 am).

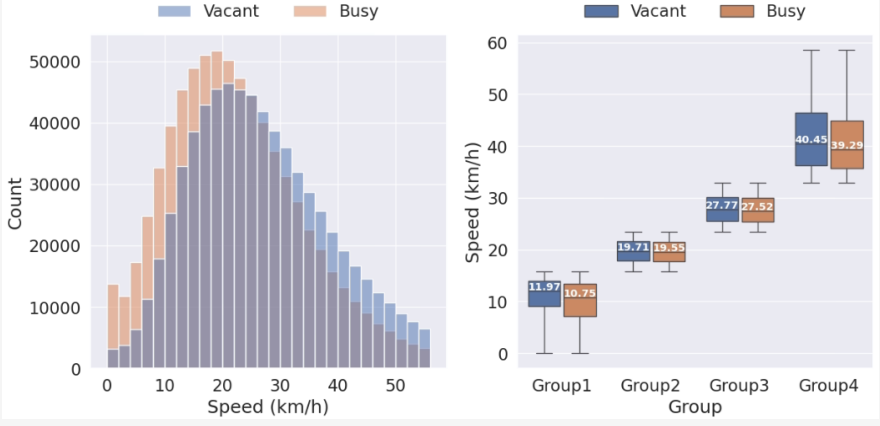

Figure 13: Distribution of vacant and busy trip speeds

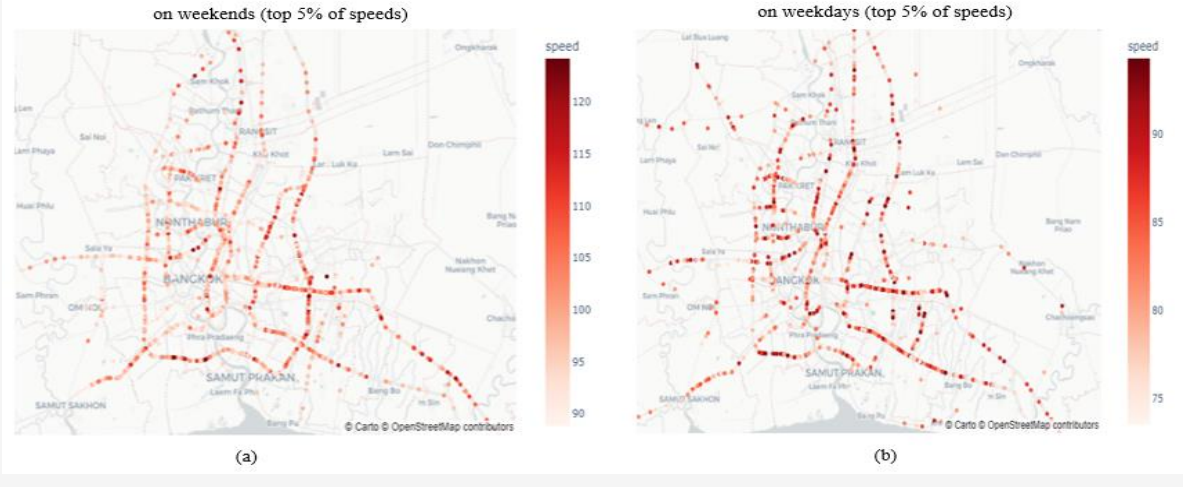

Figure 14: Distribution of taxi speeds on (a) weekends and (b) weekdays (top 5% of speeds)

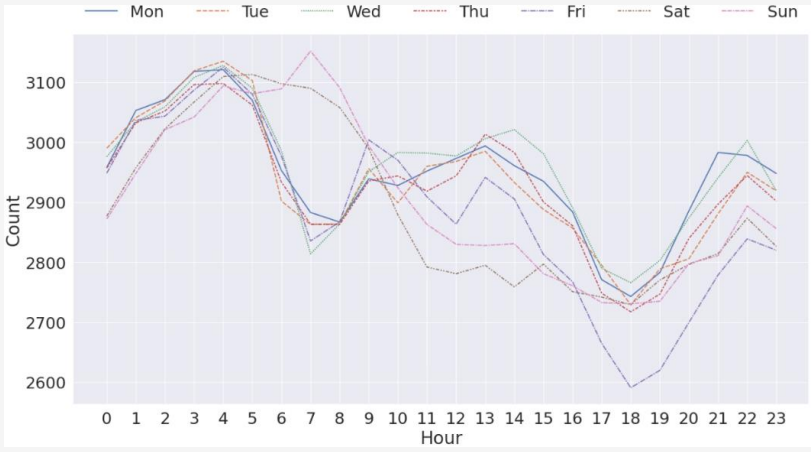

Figure 15: Average hourly trip volume of the stay period each day of the week

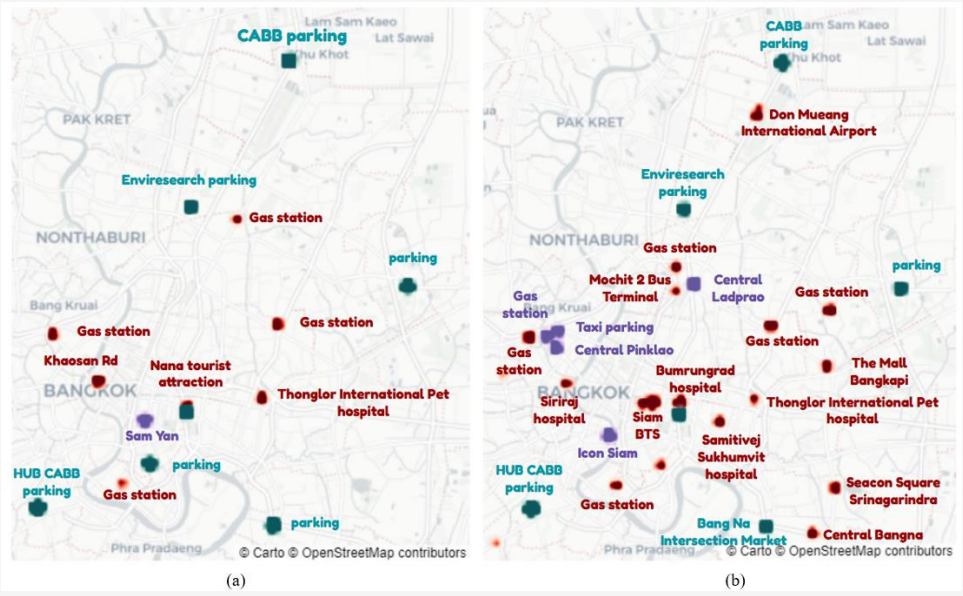

Figure 16: Density map of stay durations lasting 0–15 min (red), 15–200 min (purple), and >200 min (green) during (a) 12–1 am and (b) 6–7 am

Stay durations in hotspots can be divided into 0–15, 15–200, and >200 min stay periods. We examined the stay hotspots during midnight (12–1 am) and the morning (6-7 am). In the midnight period (Figure 16(a)), most taxis remained parked for <15 min at tourist attractions, such as Khaosan and Nana, and gas stations. The places where taxis parked for >200 min were parking hubs scattered on the outskirts of the city center. Sam Yan, a food area near a university, was the only place with a mid-range stay duration (15–200 min). Short-duration (<15 min) stays were more common during the morning than during the night (Figure 16(b)). Such stays often occurred at transport hubs, such as Mochit 2, Don Mueang Airport, and the Siam BTS station, shopping malls, and hospitals.

Similarly, taxi stationary periods of around 15–200 min more often occurred at night at places similar to those of short-stay periods, but at suburban locations.

3.6 Processing Performance on Apache Spark

We evaluated the Spark performance through trip reconstruction on probe taxi datasets of different sizes: one day (1,897,740 records), ten days (19,962,522 records), and one month (57,714,037 records). Each dataset was tested using different Spark configurations, including instance number, executor core, and executor memory. The Spark configuration consisted of five combinations, as listed in Table 3. Figure 17 illustrates the execution times for the different configurations and data sizes.

Table 3: Trip creation process times for the different Spark configurations

Instance	Core	Memory	Process time [min]					
			1 day	% decrease	10 days	% decrease	1 month	% decrease
2	1	4 GiB	2.17	baseline	22.66	baseline	55.79	baseline
2	2	4 GiB	1.28	41%	15.87	30%	30.08	46%
3	2	4 GiB	0.98	55%	7.05	69%	26.77	52%
2	4	8 GiB	0.82	62%	6.22	73%	22.23	60%
3	3	6 GiB	0.83	62%	5.2	77%	17.51	69%

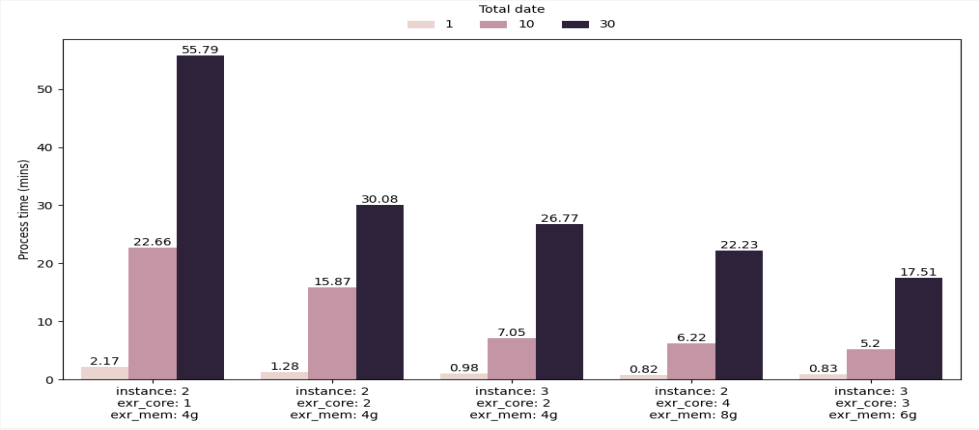

Figure 17: Spark execution times for different configurations and data sizes

After generating the trips, the output sizes of the one-day, ten-day, and one-month datasets were 70,549, 769,448, and 2,234,344 records, respectively, accounting for approximately 3.8% of the input data. The execution time rapidly decreased when the number of execution cores was adjusted, particularly with the one-month dataset, in which the execution time decreased by approximately 46%. Although increasing the number of instances or cores would improve the performance, the execution time may no longer decline linearly when the increase reaches a certain point.

4. Conclusion

In this study, we proposed a trip analysis toolkit approach for constructing trips using taxi trajectory data and conducted a mobility analysis. We identified stationary periods using a distance and duration threshold to indicate which points are stop points. Based on this, we then defined moving segments along the intervals between the stay segments. Each moving segment is identified as either busy or vacant. Furthermore, we selected a busy segment to create an origin and destination. Finally, trip and OD data were used to analyze mobility indicators, including travel patterns, accessibility, speed, and stay spots. Although this trip construction method specifically used taxi probe data, the approach can be applied to similarly structured data by modifying the appropriate key thresholds in the trip creation process. The software was developed using the Python and PySpark libraries, which allowed the distributed processing and support of large-scale data. We demonstrated the use of the software toolkits by analyzing real-world taxi data in Bangkok, Thailand using a set of indicators supported by the software. The performance measurements for large-scale processing were evaluated and the usefulness of the toolkit was proven. This approach can serve as a foundational software for analyzing vehicle-based trajectory data using predefined indicators. It's important to note that this toolkit currently does not analyze data based on road networks.

However, a road-based map matching technique can be utilized to gather information about road segments. Despite this limitation, the indicators and intermediate outputs produced by the toolkit can be leveraged to enhance taxi services, improve the work quality of taxi drivers, and increase their earnings. Furthermore, it can deliver substantial benefits for fleet management and fuel efficiency.

Acknowledgments
This study was partially supported by the Sirindhorn International Institute of Technology (SIIT), Thammasart University.

References

[1] Guan, X., Chen, C., Ren, I., Yeung, K. Y., Hung, L. H. and Lloyd, W. J., (2022). Mobility Analysis Workflow (MAW): An Accessible, Interoperable, and Reproducible Container System for Processing Raw Mobile Data. *arXiv:2204.09125*. http://arxiv.org/abs/2204.09125.

[2] Amoroso, S., Caruso, L. and Castelluccio, F., (2011). Indicators for Sustainable Mobility in the Cities. *WIT Transactions on Ecology and the Environment*, Vol. 148. https://doi.org/10.2495/RAV110241.

[3] Peungnumsai, A., Witayangkurn, A., Nagai, M., Arai, A., Ranjit, S. and Ghimire, B. R., (2017). Bangkok Taxi Service Behavior Analysis Using Taxi Probe Data and Questionnaire Survey. *Proceedings of the 4th Multidisciplinary International Social Networks Conference (MISNC '17), New York, NY, USA, July 17–19, 2017*, New York: Association for Computing Machinery, 2017. 1–8. https://doi.org/10.1145/3092090.3092117.

[4] Panichpapiboon, S. and Khunsri, K., (2022). A Big Data Analysis on Urban Mobility: Case of Bangkok. *IEEE Access*, Vol. 10. https://doi.org/10.1109/ACCESS.2022.3170068.

[5] Khunsri, K. and Panichpapiboon, S., (2021). A Big Data Analysis on Efficiency of Bangkok Taxi System. *2021 18th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand, May 19–22, 2021*, 39–42. https://doi.org/10.1109/ECTI-CON51831.2021 .9454833.

[6] Department of Land Transport of Thailand, (2024). Other Information, Complaint Handling, the Fiscal Year 2024. [Online]. Available: https://web.dlt.go.th/statistics/. [Accessed Aug. 10, 2024].

[7] Witayangkurn, A., Horanont, T., Ono, N. and Sekimoto, Y., (2013). Trip Reconstruction and Transportation Mode Extraction on Low Data Rate GPS Data from Mobile Phone. *Proceedings of the International Conference on Computers in Urban Planning and Urban Management, CUPUM 2013, Utrecht, Netherlands*, 1-19.

[8] Apache Software Foundation. Hadoop. [Online]. Available: https://hadoop.apache.org. [Accessed Aug. 1, 2024].

[9] Apache Software Foundation. Spark. Available: [Online]. https://spark.apache.org [Accessed Aug. 1, 2024].

[10] Ibtisum, S., Bazgir, E., Rahman, S. A. and Hossain, S. S., (2023). A Comparative Analysis of Big Data Processing Paradigms: Mapreduce vs. Apache Spark. *World Journal of Advanced Research and Reviews*, Vol. 20. https://doi.org/10.30574/wjarr.2023.20.1.2174.

[11] Vidović, K., Šoštarić, M. and Budimir, D., (2019). An Overview of Indicators and Indices Used for Urban Mobility Assessment. *Promet - Traffic and Transportation*, Vol. 31. https://doi.org/10.7307/ptt.v31i6.3281.

[12] Yang, Y., Jia, B., Yan, X. Y., Li, J., Yang, Z. and Gao, Z., (2022). Identifying Intercity Freight Trip Ends of Heavy Trucks from GPS Data. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 157. https://doi.org/10.1016/j.tre.2021.102590.

[13] Fu, Z., Tian, Z., Xu, Y. and Qiao, C., (2016). A Two-Step Clustering Approach to Extract Locations from Individual GPS Trajectory Data. *ISPRS International Journal of Geo-Information*, Vol. 5. https://doi.org/10.3390/ijgi5100166.

[14] Witayangkurn A., Horanont T., Nagai M. and Shibasaki R., (2018). Large Scale Mobility Analysis: Extracting Significant Places using Hadoop/Hive and Spatial Processing. *Advances in Intelligent Systems and Computing: International Conference on Knowledge, Information, and Creativity Support, KICSS 2015, Phuket, Thailand, November 12–14, 2015*, Theeramunkong T., Skulimowski A., Yuizono T., Kunifuji S., Eds. Springer, Cham. 205–219.

[15] Trip Analysis Toolkits. *GitHub*. Available: [Online]. https://github.com/SpatialDataCommons/Trip-Analysis-toolkits [Accessed Sep. 20, 2024].