Insights from a Spatiotemporal Analysis of Top-Performing Bangkok Taxis

Sooksan Panichpapiboon
School of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
sooksan@alumni.cmu.edu

Kaiden Semapakdi-Chang
United Lisbon International School
Lisboa, Portugal
kaidenksc@gmail.com

***Abstract*—Taxi is one of the most widely used public transportation systems. It provides a convenient way for people to move around a city. Nonetheless, past studies have shown that traditional street-hail taxis are highly inefficient, as most taxis spend most of their operating time without passengers aboard. However, there are still highly efficient taxis that outperform the others. Learning how these top performers stay ahead of their peers is crucial. This paper analyzes over four million taxi trips in Bangkok and its nearby region to extract meaningful insights from top-performing taxis. Spatial and temporal characteristics of trips made by top performers and ordinary taxis are investigated. These insights are valuable for improving the current taxi service and helpful in planning future mobility services such as autonomous taxis.**

***Index Terms*—Spatiotemporal analysis, big data analytics, smart mobility, smart city**

I. Introduction

Taxi is a transportation system popularly used in many major cities. However, it has been shown that traditional street-hail taxi systems have low efficiency [1]–[3]. In fact, most taxis spend most of their operating time without passengers aboard. Despite this fact, there are still highly efficient taxis that outperform the others. Learning how these top performers stay ahead of their peers is crucial. Do top performers search for passengers differently from ordinary taxis? Are there any areas where top performers commonly search for passengers? Do top performers and ordinary taxis select different hours to operate? Is it easier to search for passengers on the weekdays or the weekends? These are some of the critical research questions investigated in this study. The insights drawn from this study are valuable for improving the current taxi service and helpful in planning future mobility services such as autonomous taxis.

In this paper, we analyze approximately 4.6 million taxi trips in Bangkok and its nearby region in January, April, July, and October 2022. These trips are extracted from the real global positioning system (GPS) traces of Bangkok taxis, publicly provided by the Thai Intelligent Traffic Information Center (iTIC) [4]. Unlike most public taxi datasets, which only keep records of busy trips (i.e., trips with passengers), the Bangkok taxi dataset also provides records of taxis when they are vacant. Consequently, it enables us to genuinely observe important characteristics of vacant trips. Unlike other studies, which can only speculate on the vacant trips from secondary sources, this dataset allows us to extract the vacant trips directly from the actual trajectories. Undoubtedly, taxis that can search for passengers quickly and minimize their vacant time are more efficient than others. Having accurate vacant trip data is essential to analyzing the taxis' performance. In this study, we investigate the characteristics of trips made by the most efficient taxis (i.e., top performers) and those around the median (i.e., ordinary taxis).

The contributions of this paper can be summarized as follows:

- We demonstrate that the top performers generally outperform the ordinary taxis by a factor of two. The vacant time of the top performers is substantially less than that of the ordinary taxis. On average, the top performers can successfully seek passengers within 15 minutes, which is only half that of the ordinary taxis. This suggests that there is still plenty of room for improving the current taxi system. A policy or strategy that can help ordinary taxis perform better is still much needed.
- We show that there is a consistent spatial pattern of locations where the top performers search for passengers. This pattern is consistent across different months. Moreover, there is a consistent spatial pattern of locations where the ordinary taxis search for passengers. This pattern is also consistent across different months. However, the spatial patterns of the two taxi groups are distinctly different. These spatial patterns offer great insights into recommending the taxis on where to look for passengers to minimize their vacant time.
- We demonstrate that there is a temporal difference between the hours at which the two taxi groups prefer to operate. The best and worst periods to operate are identified. Lastly, we show that there is no significant difference between operating on the weekdays and the weekends. These are practical insights for planning the operation hours for future mobility services such as autonomous taxis.

The rest of this paper is organized as follows. In Section II, we briefly discuss related work. Data and feature extraction are explained in Section III. The analysis is discussed in

Section IV. Finally, we conclude this paper in Section V.

II. Related Work

As many sets of taxi trajectory data become more openly accessible, they are used in a variety of ways. Applications of these data include demand prediction, route recommendation, efficiency analysis, and intelligence extraction. Historical pickup data are typically used for building demand prediction models [5]–[8] and route recommendation models [9]–[12]. Since taxi trajectory data genuinely reflect how the taxis operate, they are a critical component of studies on efficiency evaluation [13]–[15]. Moreover, taxi trajectory data also allow one to learn how high-performance taxis operate. In this section, we briefly discuss work related to extracting insights from high-performance taxis.

Many studies have analyzed taxi trajectory data to extract meaningful insights from high-performance drivers. In [16], the authors observed where high-performance taxis picked up their passengers and grouped these locations into clusters of recommended hotspots. Similarly, in [17], the authors extracted high-income regions from the taxi data in Xi'an, China. In [18], the authors identified efficient operations from high-performance taxis in Changsha, China. However, the vacant trips were completely neglected. As a result, the analysis might not accurately reflect the actual efficiency. In [19], the authors show that high-performance taxis typically spend around 25 to 35 percent less searching time than low-performance taxis. However, the passenger searching time was indirectly estimated from the gap time between two passenger delivery trips, which could be inaccurate.

Unlike these existing works, in which the vacant trips were either neglected or estimated indirectly from secondary sources, we investigate the characteristics of the vacant trips directly from the actual data. Extending from our prior work [2], [3], we extract meaningful insights from top-performing and ordinary taxis in this paper.

III. Data

In this section, we describe the data and the fundamental features used in our analysis.

A. Data Description and Pre-processing

In this study, we analyze the Bangkok taxi dataset, which is publicly provided by the Thai Intelligent Traffic Information Center (iTIC) [4]. The dataset contains raw GPS records of Bangkok taxis, which were collected from each taxi approximately every one to three minutes. Each GPS record contains basic positioning data such as latitude, longitude, and timestamp. Moreover, *for-hire* light status and engine status were also collected. The for-hire light status enables us to identify whether a taxi is busy (i.e., carrying a passenger) or vacant (i.e., searching for a passenger). The engine status allows us to determine whether the vehicle engine is running. In this study, we concentrate on the taxi trips in Bangkok and its nearby region during January, April, July, and October 2022. A map of the study area is shown in Fig. 1. The geolocations of the four corners of the study area are given in Table I.

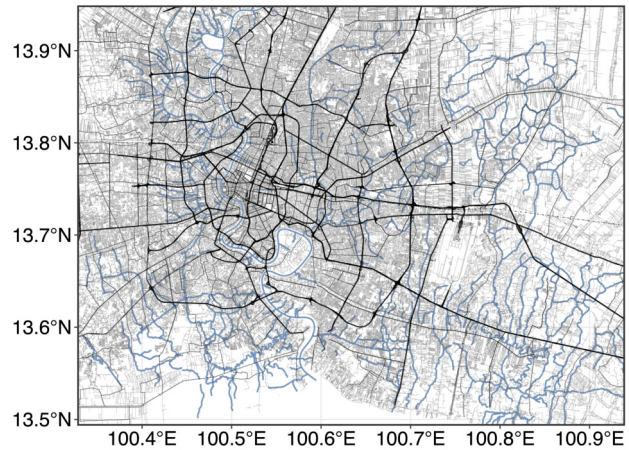

Fig. 1. The study area covers Bangkok and some parts of its nearby region. (This map is created from the open data provided by OpenStreetMap [20] under the Open Database Licence [21].)

TABLE I
Four corners of the study area.

Corner	Latitude	Longitude
Upper left	13.94913	100.32750
Upper right	13.94913	100.93963
Lower left	13.49310	100.32750
Lower right	13.49310	100.93963

The raw GPS data are pre-processed as follows. Since we only consider the period when each taxi is active (i.e., when the engine is running), all the records with inactive engine status are filtered out. In addition, all the records with invalid GPS status are also filtered out. Finally, any records with geolocations outside of the study area are excluded. After the initial pre-processing, we are left with valid records of active taxis inside the study area. Essential features are then extracted from these data, as described in the next section.

B. Feature Extraction

After the initial pre-processing, the following features are extracted from the data:

1) *Trip:* A *trip* is a sequence of consecutive GPS records with the same for-hire status. Two types of trips can be extracted. They are referred to as *vacant trips* and *busy trips*. A vacant trip or a trip without passengers is defined as a sequence of consecutive GPS records with active for-hire status. In contrast, a busy trip or a trip with passengers is defined as a sequence of consecutive GPS records with inactive for-hire status. The first GPS record in a sequence is regarded as the starting point of the trip, and the last GPS record in the sequence is regarded as the end point of the trip.
2) *Trip Duration:* Trip duration is the difference between the timestamps at the start and at the end of the trip. For

TABLE II
SUMMARY OF TAXI TRIP DATA.

Description	Jan	Apr	Jul	Oct
Taxis	3,222	3,021	3,153	3,241
Busy trips	551,263	509,332	583,527	607,189
Vacant trips	572,560	536,921	626,127	655,890
Total trips	1,123,823	1,046,253	1,209,654	1,263,079

example, suppose that trip A consists of four GPS points, namely P_1, P_2, P_3, and P_4. The duration of trip A can be calculated from the difference between the timestamps of P_1 and P_4.

3) *Trip Distance:* Trip distance is an aggregated sum of the distance between consecutive GPS points in the trip. For example, suppose that trip A consists of four GPS points, namely P_1, P_2, P_3, and P_4. Let D_{ij} be a geographical distance between P_i and P_j. Then, the total distance of trip A is estimated as $D(A) = D_{12} + D_{23} + D_{34}$.

A summary of the extracted trip data is given in Table II. In each month, more than three thousand taxis are operational. These taxis collectively made over one million trips each month.

IV. SPATIOTEMPORAL ANALYSIS

A. Top Performers and Ordinary Taxis

This study aims to investigate the differences between the characteristics of trips made by top-performing and ordinary taxis. Therefore, we must first define a performance metric for categorizing the taxis into these two groups. In this study, the performance of each taxi is measured by the average duration of its vacant trips. Certainly, a taxi that can quickly search for passengers and consistently minimize its vacant time is more efficient than others. In each month, the taxis are ranked by their average vacant time. The average vacant time of a taxi i can be calculated as

$$\overline{V_i} = \frac{1}{K_i} \sum_{k=1}^{K_i} V_{i,k} \quad (1)$$

where $V_{i,k}$ is the duration of the kth vacant trip that the taxi i made in the month, and K_i is the total number of vacant trips that the taxi i made in the month. To ensure that we investigate the taxis that operate regularly, we only consider the taxis that made at least thirty vacant trips during the month (i.e., an average of at least one trip a day).

In each month, the top one hundred taxis with the smallest average vacant time are identified. These taxis are defined as the *top performers* for the month. In contrast, ordinary taxis have medium performance. After the taxis are ordered by their average vacant time, the median of the average vacant time is determined. The one hundred taxis with the average vacant time closest to the median are designated as the *ordinary taxis*. The list of top performers and the list of ordinary taxis can vary across different months.

TABLE III
SUMMARY OF TRIPS MADE BY TOP PERFORMERS AND ORDINARY TAXIS.

Description	TOP	ORD
Busy trips	172,731	109,489
Mean trip duration (min)	36.73	37.32
Mean trip distance (km)	6.99	7.53
Vacant trips	176,479	115,534
Mean trip duration (min)	15.57	29.37
Mean trip distance (km)	7.46	10.37

A statistic of the trips made by the top performers (TOP) and the ordinary taxis (ORD), considering all four months, is shown in Table III. It can be observed that the characteristics of the busy trips in these two groups are not significantly different. The mean duration of the busy trips made by the top performers and that of the ordinary taxis differ by less than a minute. Similarly, the discrepancy between the mean distance of the busy trips made by the top performers and that of the ordinary taxis is smaller than one kilometer. On the contrary, the characteristics of the vacant trips in these two groups are distinctly noticeable. The mean vacant duration of the top performers is almost half that of the ordinary taxis. In addition, the mean distance of the vacant trips made by the top performers is approximately 30% less than that of the ordinary taxis. Since the significant differences between the two groups are from the vacant trips, we will mainly analyze the vacant trips in the following sections.

B. Spatial Analysis

In this section, we analyze the spatial characteristics of the vacant trips made by the top performers and the ordinary taxis. Learning where the taxis in these two groups search for passengers could yield significant insights into why one group outperforms the other. Fig. 2 illustrates the density of the starting points of the vacant trips made by the top performers in each month. It can be observed that the four density plots exhibit a similar pattern. There are two main clusters where the vacant trips originate. The highest intensity cluster is centered around the coordinate (13.775, 100.48), which is a busy neighborhood on the west side of Bangkok. The second highest intensity cluster is located around the coordinate (13.725, 100.53), which is a bustling business district with office buildings. The overall spread of each spatial distribution is also small. This suggests that the top performers focus their search in limited areas instead of hunting for passengers in all city areas. Since the observed pattern is consistent across all four months, it implies that these are the golden spots where the taxis should look for their passengers to minimize their vacant time.

In contrast to the case of the top performers, a noticeably different spatial distribution pattern is observed in the case of the ordinary taxis. Fig. 3 illustrates the density of the starting points of the vacant trips made by the ordinary taxis in each month. The spatial distribution looks consistent across the four months. Contrary to what is observed in the case of the top

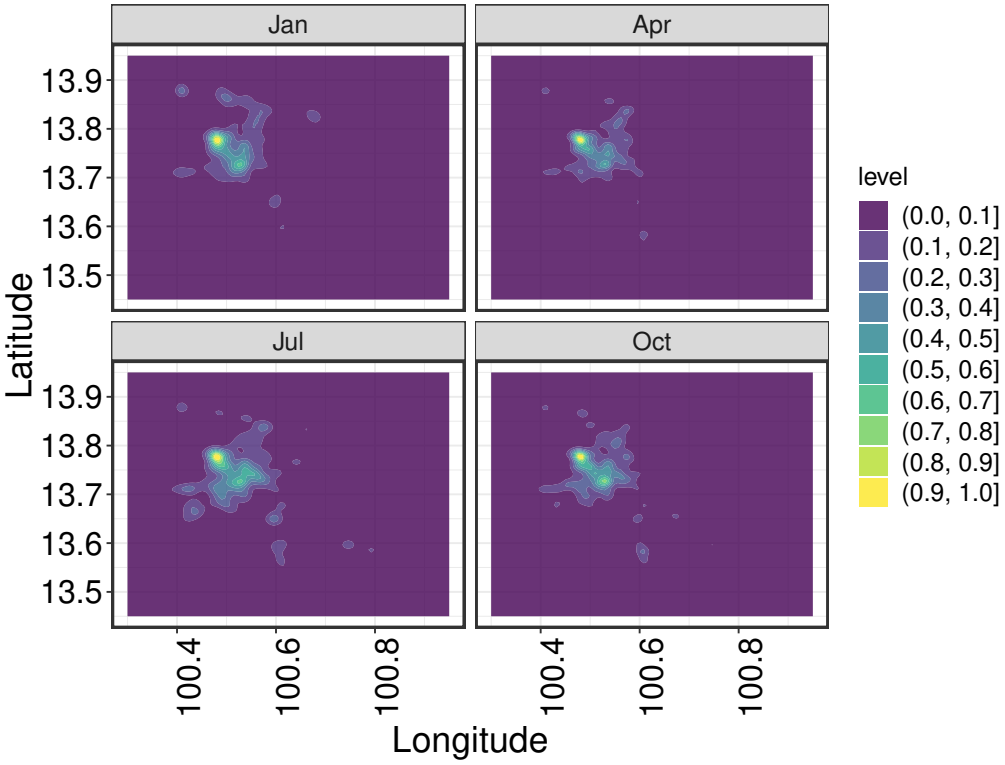

Fig. 2. Density of the starting points of vacant trips made by the top performers.

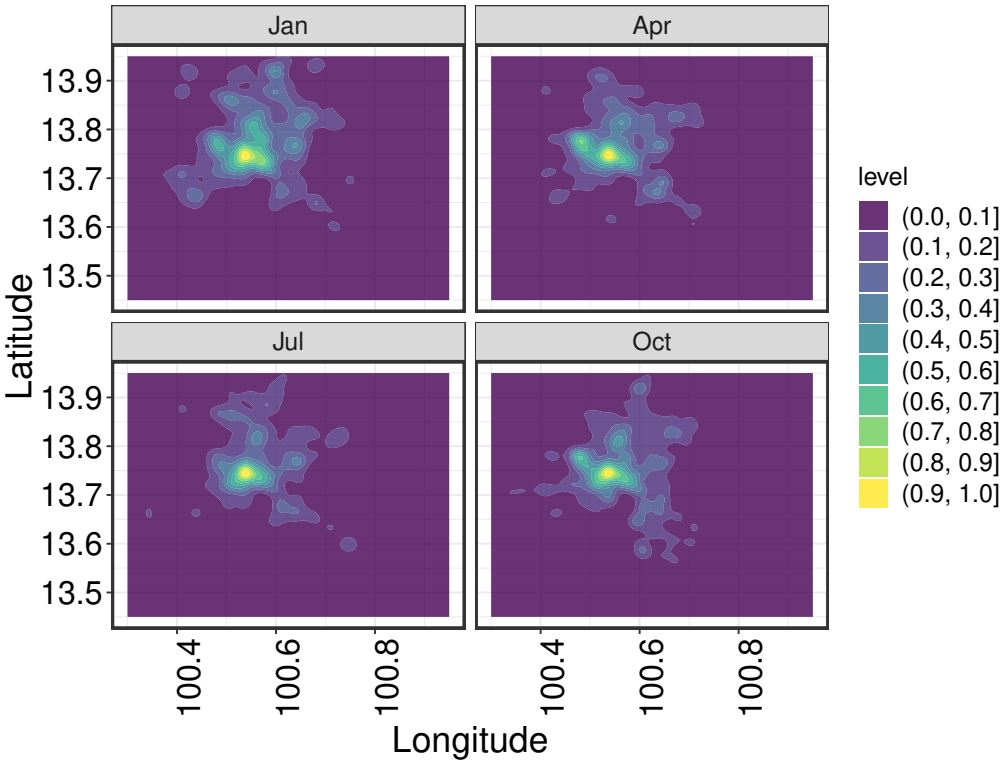

Fig. 3. Density of the starting points of vacant trips made by the ordinary taxis.

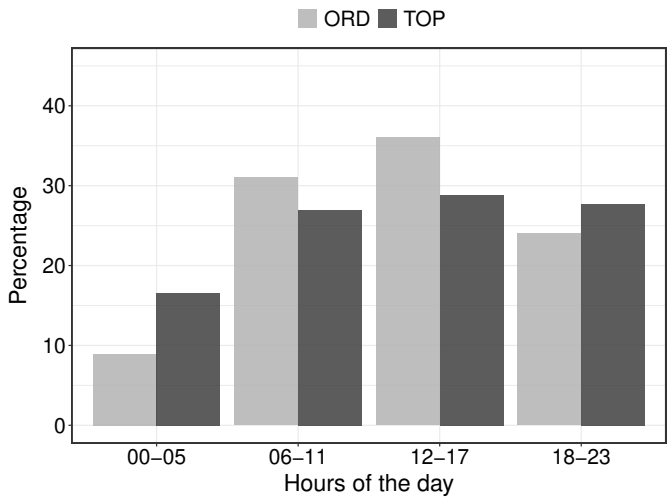

Fig. 4. Percentage of vacant trips in each period of the day.

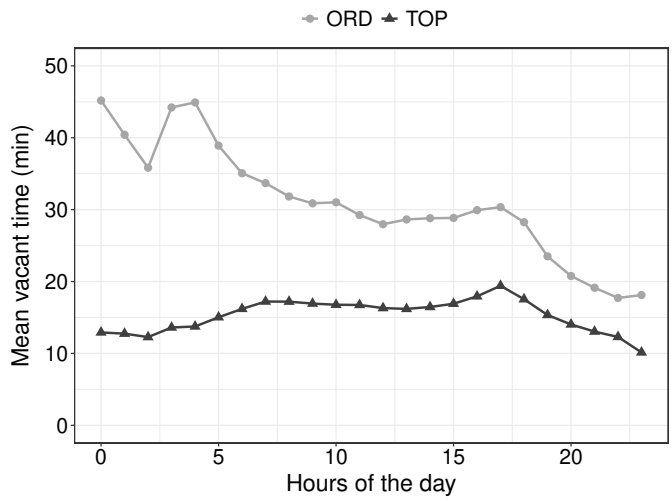

Fig. 5. Mean vacant time at each hour of the day.

performers, however, this pattern has a single major cluster surrounded by multiple minor clusters. The highest intensity cluster is located around the coordinate (13.745, 100.54), which is a vibrant district with shopping malls, restaurants, and hotels. Compared to the case of the top performers, the distribution spread is much larger. This suggests that the ordinary taxis search for passengers more dispersedly across different city areas.

C. Temporal Analysis

In this section, we analyze the temporal characteristics of the vacant trips made by the top performers and the ordinary taxis. First, we investigate if the two taxi groups have different preferences on the operation hours. The temporal distributions of the vacant trips made by the top performers and the ordinary taxis in four periods of the day during the four months are shown in Fig. 4. These periods are the early morning hours (00-05), the morning hours (06-11), the afternoon hours (12-17), and the evening hours (18-23). In the case of the top performers, the vacant trips are approximately equally distributed in the morning, afternoon, and evening hours. Around 84% of the vacant trips made by the top performers are in these three periods. In the case of the ordinary taxis, the highest percentage of trips are in the afternoon, and the second highest period is in the morning. Approximately 67% of the vacant trips made by the ordinary taxis are in these two periods.

Next, we investigate how the performance of the taxis varies across different hours of the day during the four months. Fig. 5 illustrates the mean vacant time of the two taxi groups at each hour of the day. The mean vacant time of the top performers ranges from around 10 to 19 minutes. Surprisingly, the mean vacant time during the daytime (i.e., from 6 am to 6 pm) is larger than that of the nighttime. Apparently, it takes the top performers more time to find passengers during the daytime. In contrast to the case of the top performers, the mean vacant time of the ordinary taxis has a much larger variation. It ranges from around 18 to 45 minutes. The ordinary taxis perform best in the evening from 7 pm to 11 pm, whereas the worst hours are in the early morning from 12 am to 5 am. In terms of hourly mean vacant time, the top performers outperform the ordinary taxis by a factor of two, on average.

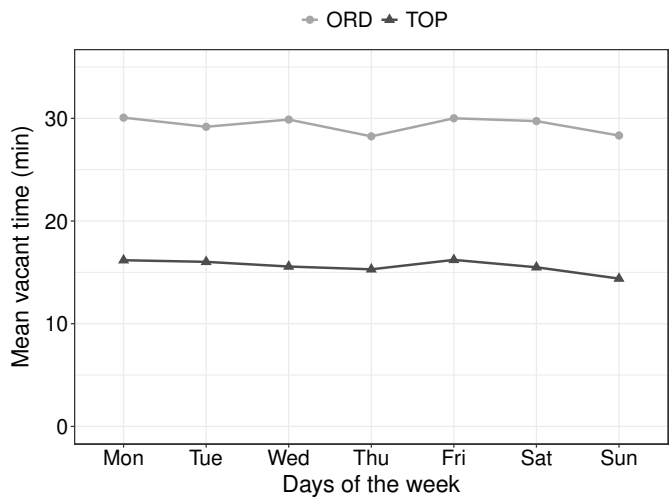

Fig. 6. Mean vacant time in each day of the week.

Finally, we investigate if there is a significant difference between trips on the weekdays and those on the weekends. Fig. 6 illustrates the mean vacant time across different days of the week during the four months. It can be observed that the mean vacant time is relatively stable across different days of the week. There is no significant discrepancy between the weekday trips and the weekend trips. This is reflected in both taxi groups. Once again, we notice that the top performers outperform the ordinary taxis by a factor of two.

V. Conclusion

Based on the real trajectory traces of Bangkok taxis, we analyze the characteristics of the trips made by the top performers and those made by the ordinary taxis. Unlike most existing studies, which either neglected or speculated on the vacant trips from secondary sources, this study extracts the vacant trips directly from the actual data. Learning what the top performers do differently from the ordinary taxis provides great insights into a good operational strategy. Here are the valuable insights obtained from this study:

- Surprisingly, the mean duration and the mean distance of the busy trips made by the top performers are not significantly different from those of the ordinary taxis. Their mean busy durations differ by less than a minute, while their mean distances differ by less than a kilometer. A policy that aims to improve the duration and distance of a busy trip would not be impactful because the ordinary taxis have already performed as well as the top performers in this aspect. The room for improvement in terms of the busy trips is small.
- In contrast, the vacant time or the passenger searching time of the top performers is substantially different from that of the ordinary taxis. On average, the time it takes the top performers to successfully seek new passengers is only half of that required by the ordinary taxis. In order to make a significant improvement, the strategy should concentrate on reducing the vacant time of the ordinary taxis. This could be done by recommending them to search for passengers in the right area at the right time.
- Our spatial analysis reveals that the top performers and the ordinary taxis search for passengers at significantly different locations. The spatial distributions of the vacant trips made by the top performers and those made by the ordinary taxis are distinctly different. Compared to the ordinary taxis, the top performers concentrate their search in much more limited areas. Focusing on these areas leads to a substantially smaller search time.
- Finally, our temporal analysis reveals that the top performers and the ordinary taxis have different preferences in operation hours. The top performers equally prefer to operate in the morning, afternoon, and evening hours. In contrast, the ordinary taxis prefer to operate in the afternoon hours. Moreover, compared to the ordinary taxis, the mean vacant time of the top performers has less variation across the hours of the day. Lastly, within each taxi group, there is no significant difference between weekday and weekend trips. The average vacant trip times are similar across different days of the week.

These insights are beneficial to improving the current taxi systems and valuable to planning future mobility services such as autonomous taxis.

References

[1] Y. Chen, M. Hyland, M. P. Wilbur, and H. S. Mahmassani, "Characterization of taxi fleet operational networks and vehicle efficiency: Chicago case study," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 48, pp. 127–138, Oct. 2018.

[2] K. Khunsri and S. Panichpapiboon, "A big data analysis on efficiency of Bangkok taxi system," in *Proc. 18th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, May 2021, pp. 39–42.

[3] S. Panichpapiboon and K. Khunsri, "A big data analysis on urban mobility: Case of Bangkok," *IEEE Access*, vol. 10, pp. 44400–44412, 2022.

[4] Intelligent Traffic Information Center (iTIC) Foundation (2021). *Historical raw vehicles and mobile probes data in Thailand.* [Online]. Available: https://org.iticfoundation.org/download

[5] M. S. Jamil and S. Akbar, "Taxi passenger hotspot prediction using automatic ARIMA model," in *Proc. 3rd Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2017, pp. 23–28.

[6] K. Zhang, Z. Feng, S. Chen, K. Huang, and G. Wang, "A framework for passengers demand prediction and recommendation," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Jun. 2016, pp. 340–347.

[7] J. Xu, R. Rahmatizadeh, L. Bölöni, and D. Turgut, "Real-time prediction of taxi demand using recurrent neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2572–2581, Aug. 2018.

[8] K. Niu, C. Cheng, J. Chang, H. Zhang, and T. Zhou, "Real-time taxi-passenger prediction with L-CNN," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4122–4129, May 2019.

[9] B. Qu, W. Yang, G. Cui, and X. Wang, "Profitable taxi travel route recommendation based on big taxi trajectory data," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 653–668, Feb. 2020.

[10] F. Zong, T. Wu, and H. Jia, "Taxi drivers' cruising patterns–insights from taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 2, pp. 571–582, Feb. 2019.

[11] R. Wang, C.-Y. Chow, Y. Lyu, V. C. S. Lee, S. Kwong, Y. Li, and J. Zeng, "TaxiRec: recommending road clusters to taxi drivers using ranking-based extreme learning machines," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 585–598, Mar. 2018.

[12] T. Wang, Z. Shen, Y. Cao, X. Xu, and H. Gong, "Taxi-cruising recommendation via real-time information and historical trajectory data," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 7898–7910, Aug. 2023.

[13] W. Zhai, X. Bai, Z. Peng, and C. Gu, "A bottom-up transportation network efficiency measuring approach: A case study of taxi efficiency in New York City," *J. Transp. Geography*, vol. 80, p. 102502, Oct. 2019.

[14] X. Zhan, X. Qian, and S. V. Ukkusuri, "A graph-based approach to measuring the efficiency of an urban taxi service system," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2479–2489, Sep. 2016.

[15] X. Dong, M. Zhang, S. Zhang, X. Shen, and B. Hu, "The analysis of urban taxi operation efficiency based on GPS trajectory big data," *Phys. A, Stat. Mech. Appl.*, vol. 528, p. 121456, Aug. 2019.

[16] Y. Ge, C. Liu, H. Xiong, and J. Chen, "A taxi business intelligence system," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 735–738.

[17] C. Yuan, X. Geng, and X. Mao, "Taxi high-income region recommendation and spatial correlation analysis," *IEEE Access*, vol. 8, pp. 139529–139545, 2020.

[18] H. Rong, Z. Wang, H. Zheng, C. Hu, L. Peng, Z. Ai, and A. K. Sangaiah, "Mining efficient taxi operation strategies from large scale geo-location data," *IEEE Access*, vol. 5, pp. 25623 – 25634, 2017.

[19] X. Zhou, H. Rong, C. Yang, Q. Zhang, A. V. Khezerlou, H. Zheng, Z. Shafiq, and A. X. Liu, "Optimizing taxi driver profit efficiency: A spatial network-based Markov decision process approach," *IEEE Trans. Big Data*, vol. 6, no. 1, pp. 145–158, Mar. 2020.

[20] (2023). *OpenStreetMap*. Accessed: Mar. 7, 2023. [Online]. Available: https://www.openstreetmap.org.

[21] (2023). *Open Data Commons*. Accessed: Mar. 7, 2023. [Online]. Available: https://opendatacommons.org.