Insights from Spatiotemporal Analysis of Taxi Supply-Demand Imbalance in Bangkok

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*—Supply-demand imbalance is a critical problem in street-hail taxi systems. In a high-demand or overserved area, where pickups exceed drop-offs, taxis have to compete intensely for passengers. This results in a reduction of their revenues. On the other hand, in a low-demand or underserved area, where taxis are less likely to search for passengers, the passengers experience a longer waiting time. Moreover, taxis often perceive that finding passengers in a low-demand area is more difficult. This results in the refusal of service for trips to destinations in low-demand areas, which is one of the critical problems in the Bangkok taxi service. In this study, based on more than four million real taxi trips in Bangkok, we quantify the effects of the supply-demand imbalance on passenger search time and distance. In contrast to the conventional perception, we demonstrate that, on average, taxis in underserved areas do not have longer search times and distances than those in overserved areas. In fact, there is almost no correlation between the level of supply-demand imbalance and passenger search time and distance. These new insights change how recommendation and rebalancing models for traditional street-hail taxi services and future autonomous taxi services should be designed.**

***Index Terms*—Taxi supply-demand imbalance, smart mobility, smart city, big data analytics**

I. Introduction

The imbalance between supply and demand is a persistent problem in street-hail taxi systems. It causes inefficiencies that affect passengers, drivers, and city traffic. Typically, the imbalance is measured by the ratio of pickups and drop-offs. In a high-demand or overserved area, where the number of pickups exceeds the number of drop-offs, taxis have to compete intensely for passengers. This likely leads to reduced revenues. On the other hand, in a low-demand or underserved area, where taxis are less likely to search for passengers, the passengers experience a longer waiting time. Moreover, it is generally perceived that passenger search time in underserved areas is higher than in overserved areas. In other words, taxis often believe that it takes them longer to find the next passenger in the low-demand area than in the high-demand area. This results in the refusal of service for trips to destinations in the low-demand areas, which is one of the critical problems in the Bangkok taxi service.

In this study, we quantify the effects of supply-demand imbalance on passenger search time and distance. Does it take longer to find the next passengers in the underserved areas? Do taxis have to drive farther in the underserved areas to find their next passengers? Is it advisable to recommend taxis to the high-demand areas? Are most areas in Bangkok underserved or overserved? These are the key research questions that we aim to address in this study.

In this study, we analyze over four million taxi trips in Bangkok and its surrounding 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) [1]. 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, such as vacant time and distance.

The contributions of this paper can be summarized as follows:

- In contrast to the conventional belief, we demonstrate that taxis in underserved areas do not have larger passenger search times than those in overserved areas, and vice versa. Surprisingly, we find that there is almost no correlation between the passenger search time and the level of supply-demand imbalance.
- Similarly, we find that taxis in underserved areas do not drive farther to find their next passengers than those in overserved areas, and vice versa. In fact, we demonstrate that there is almost no correlation between the passenger search distance and the level of supply-demand imbalance.
- Finally, we illustrate how the level of imbalance in each city grid in Bangkok changes spatially and temporally. The dynamics of the imbalance level in each city grid can be used to infer its functional use (e.g., residential, commercial, airports, etc.), benefiting future prediction models.

These new insights change how the recommendation and rebalancing models should be designed, benefiting both traditional street-hail taxi and future autonomous taxi services.

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

Demand and supply of street-hail taxis have been studied extensively in the literature. On the demand side, most studies mainly focus on demand prediction models, which are typically created from historical pickup data. Common prediction techniques include classical time-series models, such as autoregressive integrated moving average [2]–[4], and deep learning models, such as convolutional neural networks and long short-term memory (LSTM) [5]–[7].

On the supply side, most studies focus on recommendation models, optimal search strategy modeling, and analysis of search behavior. The common goal of recommendation models is to develop algorithms that help vacant taxis find their next passengers quickly. Typically, a recommendation model suggests a few attractive hotspots where a taxi is most likely to find its next passenger [8]–[11]. The hotspots are typically ranked based on criteria such as predicted demand, estimated earnings, and expected traffic conditions. Many studies investigate the optimal passenger search strategy. The goal is to create a mathematical model that generates an optimal search route based on specific criteria such as maximizing revenue or minimizing search time. A common technique employed in these works is the Markov Decision Process (MDP) [12]–[14]. Finally, many supply-side studies focus on analyzing search behavior, which examines passenger search behaviors that lead to positive outcomes, such as higher earnings. It usually involves identifying how high-performing taxis search for passengers [15]–[18].

Although the demand and supply of street-hail taxis have been studied extensively, none of the existing work has quantitatively analyzed the impact of the imbalance on the passenger search time and distance. Through big data analytics, we address this research gap by demonstrating how the supply-demand imbalance affects passenger search time and distance.

III. Methodology

A. Data and Pre-processing

In this study, we analyze the Bangkok taxi dataset, which is publicly provided by the Thai Intelligent Traffic Information Center (iTIC) [1]. 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, including 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 focus on the taxi trips in Bangkok and its surrounding 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.

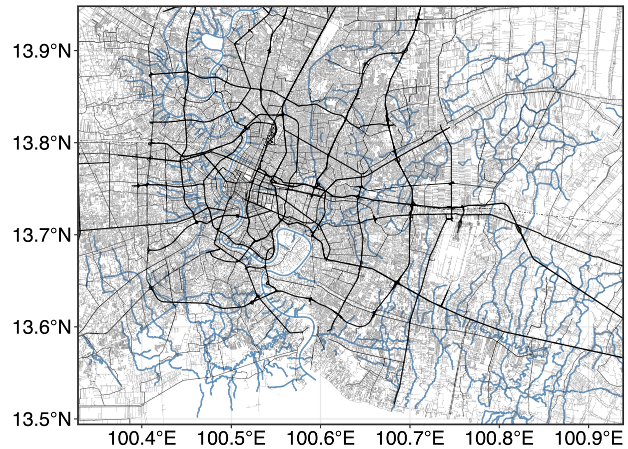

Fig. 1. The study area covers Bangkok and some parts of its nearby region. (This map is created from the open data provided by OpenStreetMap [19] under the Open Database Licence [20].)

TABLE I
FOUR CORNERS OF THE STUDY AREA.

Corner	Latitude	Longitude
Upper left	13.94500	100.32750
Upper right	13.94500	100.93833
Lower left	13.49310	100.32750
Lower right	13.49310	100.93833

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 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. Trip Extraction

After the initial pre-processing, taxi trips can be extracted. A *trip* is defined as 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.

For each trip, we can extract its duration and distance. The trip duration is the difference between the timestamps at the start and end of the trip. For 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

TABLE II
SUMMARY OF EXTRACTED TAXI TRIP DATA IN 2022.

Description	Jan	Apr	Jul	Oct
Taxis	2,874	2,657	2,754	2,826
Busy trips	520,768	482,299	553,390	575,768
Vacant trips	545,783	512,095	596,796	625,472
Total trips	1,066,551	994,394	1,150,186	1,201,240

between the timestamps of p_1 and p_4. The 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 the geographical distance between p_i and p_j. Then, the total distance of trip A is estimated as $d(A) = d_{12} + d_{23} + d_{34}$.

Additional filtering is performed to exclude obviously anomalous trips. It is extremely unlikely that a passenger will spend more than four hours in a taxi. Similarly, it is highly unlikely that a taxi will drive continuously for more than four hours to search for passengers without taking a break. As a result, trips longer than four hours are considered erroneous. A typical driving distance from one corner of the study area to its diagonally opposite corner is around 120 km. Therefore, trips farther than 160 km (i.e., 33% larger than the typical longest driving distance) will also be considered unusual. A summary of the trip data after the filtering is given in Table II. Approximately one million trips were made by these taxis each month.

C. Measurement of Supply-Demand Imbalance

The imbalance of supply and demand is measured by a metric called the *pickup-to-drop-off ratio (PDR)*. The PDR is calculated from the ratio of the number of pickups to the number of drop-offs in an observed space-time unit. More formally, the PDR in an observed space s during period t can be expressed as

$$\mathrm{PDR}(s,t) \quad = \quad \frac{P(s,t)}{D(s,t)} \tag{1}$$

where $P(s,t)$ and $D(s,t)$ are the number of pickups and the number of drop-offs in the observed space-time unit (s,t), respectively. A PDR value greater than 1 indicates that the considered space-time unit has a higher number of pickups than drop-offs; whereas, a PDR value smaller than 1 implies that the considered space-time unit has a higher number of drop-offs than pickups. Finally, the space-time unit has a perfect supply-demand balance if its PDR value is exactly equal to 1.

In this study, each observed space is defined as a 2 km $\times$ 2 km square grid, and each observed time period is one hour. With this grid size, the entire study area shown in Fig. 1 can be divided into 825 grids with 25 rows and 33 columns. Since there are 24 hours in a day, this results in a total of $825 \times 24 = 19,800$ space-time units per day. Because it is not meaningful to analyze the space-time units where both the number of pickups and drop-offs are zero, these space-time units are excluded from our consideration. The PDR values for the remaining space-time units are then calculated. However, in a space-time unit with non-zero pickups and zero drop-offs, the PDR value becomes unmeasurable (i.e., division by zero). To circumvent this numerical problem, a pseudo-count of 1 is added to both the number of pickups and the number of drop-offs in all considered space-time units. For example, a particular space-time unit might have 20 pickups and zero drop-offs. After adding the pseudo-count, the PDR ratio is computed as $(20+1)/(0+1) = 21$.

The mean PDR of each space-time unit is then computed for further analysis. In this study, we focus only on weekday trips, as most taxi trips occur on weekdays. For each space-time unit (s,t), where $s \in \{1, 2, \ldots, 825\}$ is the grid identification number and $t \in \{0, 1, \ldots, 23\}$ is the hour of the day, the mean PDR is obtained by averaging over the PDR in the unit across all the weekdays in the month. For example, the mean PDR of the space-time unit $(100, 8)$ in July 2022 is obtained by averaging over the PDR values observed in Grid 100 during the 8$^{\text{th}}$ hour, over the 21 weekdays in July 2022.

IV. SUPPLY-DEMAND IMBALANCE ANALYSIS

A. Impact of Imbalance on Search Time and Distance

In this section, we analyze the characteristics of passenger search time and distance to examine how they change with varying levels of supply-demand imbalance. For analytical purposes, we divide the space-time units into the following three categories based on their mean PDR values:

- **Underserved units**–These are low-demand units with fewer pickups than drop-offs. An underserved unit is defined as a space-time unit with a mean PDR value smaller than 0.95.
- **Balanced units**–These are units with approximately equal numbers of pickups and drop-offs. A balanced unit is defined as a space-time unit with a mean PDR value between 0.95 and 1.05.
- **Overserved units**–These are high-demand units with a higher number of pickups than drop-offs. An overserved unit is defined as a space-time unit with a mean PDR value greater than 1.05.

A box plot shown in Fig. 2 illustrates the mean vacant time of the vacant trips originating in the underserved, balanced, and overserved space-time units in July 2022. It can be observed that the characteristics of the mean vacant time in these three groups are similar. There is no substantial evidence to suggest that the vacant trips originating in an underserved unit have significantly longer duration than those originating in an overserved unit, and vice versa. The medians of the mean vacant time in the underserved, balanced, and overserved units are 22.68 minutes, 23.79 minutes, and 24.13 minutes, respectively. Thus, we observe only a slight increase in the median as the mean PDR increases. Similar characteristics are also observed in January, April, and October 2022; however, results are omitted due to space limitations.

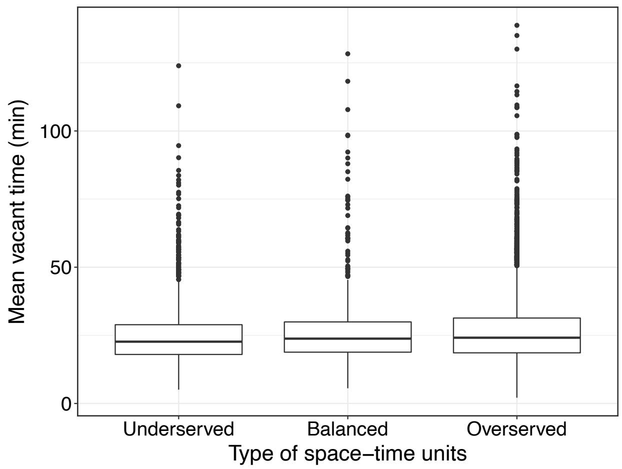

Fig. 2. A box plot of the mean vacant time of trips originating in the underserved, balanced, and overserved space-time units in July 2022.

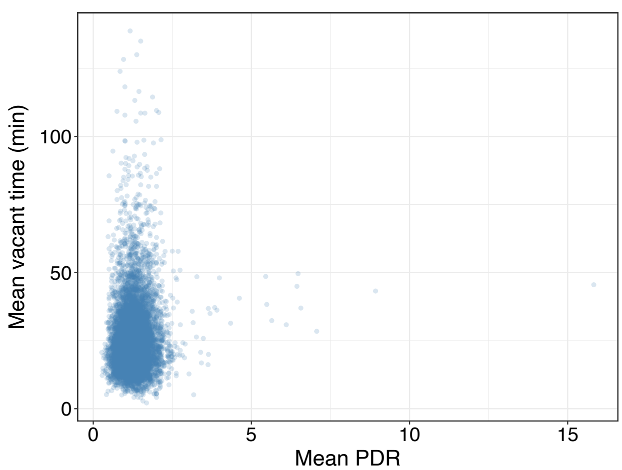

Fig. 3. Mean vacant time of trips originating in each space-time unit and its corresponding mean PDR. The data presented on this plot are from July 2022.

Fig. 3 illustrates the mean vacant time of the vacant trips originating in each space-time unit and its corresponding mean PDR in July 2022. Clearly, there is no strong correlation between the mean vacant time and the mean PDR. The Pearson correlation coefficient between these two variables is only 0.07, which confirms that they are almost uncorrelated. Surprisingly, this suggests that the vacant trips originating from an underserved unit do not have a longer duration than those from an overserved unit. In other words, taxis in low-demand areas do not have longer passenger search times than those in high-demand areas. In fact, there is almost no correlation between the type of areas and the duration of search time. This observation is consistent across the four months investigated in 2022, as shown in the second column of Table III.

Next, we investigate if there is a relation between the passenger search distance and the mean PDR. In other words,

TABLE III
CORRELATION BETWEEN MEAN PDR, MEAN VACANT TIME, AND MEAN VACANT DISTANCE.

Month	Correlation Coefficient	
	Mean PDR vs Mean Vacant Time	**Mean PDR vs Mean Vacant Distance**
Jan	0.08	0.08
Apr	0.08	0.09
Jul	0.07	0.09
Oct	0.10	0.11

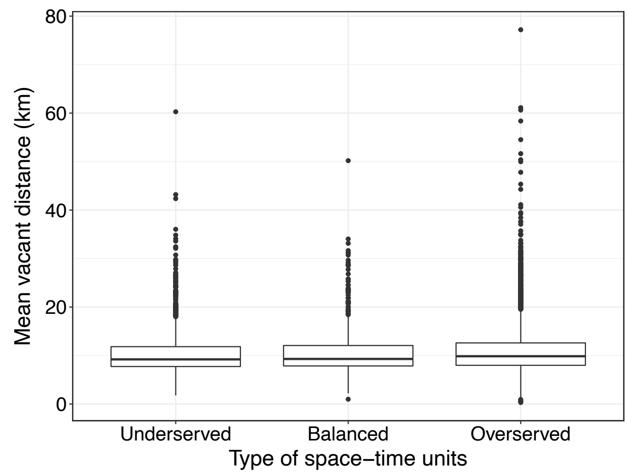

Fig. 4. A box plot of the mean vacant distance of trips originating in the underserved, balanced, and overserved space-time units in July 2022.

we analyze whether taxis in underserved areas need to drive farther to find their next passengers compared to those in overserved areas. A box plot shown in Fig. 4 illustrates the mean vacant distance of the vacant trips originating in the underserved, balanced, and overserved space-time units in July 2022. We observe that the characteristics of the mean vacant distance in these three groups are not significantly different. There is no substantial evidence to suggest that the vacant trips originating in an underserved unit have significantly larger distances than those originating in an overserved unit, and vice versa. The medians of the mean vacant distance in the underserved, balanced, and overserved space-time units are 9.21 km, 9.30 km, and 9.86 km, respectively. Apparently, the median of the vacant distance slightly increases as the mean PDR increases. Similar characteristics are also observed in January, April, and October 2022. Results are omitted due to space limitations.

Fig. 5 illustrates the mean vacant distance of the vacant trips originating in each space-time unit and its corresponding mean PDR in July 2022. Similar to what was observed earlier in the case of the mean vacant time, there is no strong correlation between the mean vacant distance and the mean PDR either. The Pearson correlation coefficient between these two variables is only 0.09, indicating that they are almost uncorrelated. This implies that the vacant trips originating in

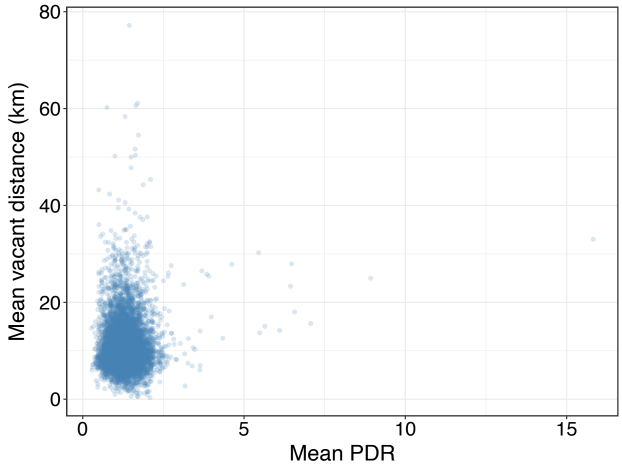

Fig. 5. Mean vacant distance of trips originating in each space-time unit and its corresponding mean PDR. The data presented on this plot are from July 2022.

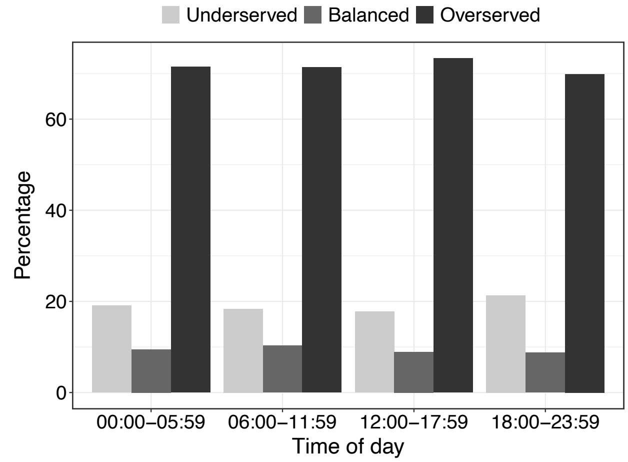

Fig. 6. Percentage of underserved, balanced, and overserved space-time units in four periods of day in July 2022.

an underserved unit do not have a longer distance than those in an overserved unit. In other words, taxis in low-demand areas do not have a longer passenger search distance than those in high-demand areas. In fact, there is almost no correlation between the type of areas and the passenger search distance. The Pearson correlation coefficients shown in the last column of Table III confirm that this observation is consistent across all four months considered in 2022.

B. Spatiotemporal Variation of Imbalance

In this section, we analyze the spatial and temporal variations of the mean PDR values. Fig. 6 illustrates the percentage of underserved, balanced, and overserved space-time units in four periods of the day in July 2022. It can be observed that the proportion of each space-time unit type is similar across the four periods of the day. In each period, approximately 20% of the units are underserved, 10% are balanced, and 70% are overserved. Clearly, the Bangkok taxi supply-demand is heavily imbalanced, with mostly overserved units. Similar characteristics are also observed in January, April, and October; however, results are omitted due to space limitations.

Next, we investigate how the mean PDR in each grid changes with the time of day. Fig. 7 illustrates the mean PDR in four periods of the day across all grids. Recall that the study area consists of 825 grids with 25 rows and 33 columns. We can observe that the mean PDR of a grid cell may change over time. For example, a grid cell can be an overserved unit in the morning and turn into an underserved unit in the evening. This is likely a characteristic of a residential area where the number of pickups is expected to be larger than the number of drop-offs in the morning, and vice versa in the evening. This suggests that the functionality of a grid cell could be inferred based on its PDR value. The temporal variation of the mean PDR in a particular area could help predict its functionality (e.g., residential, commercial, etc.).

Finally, two grid cells distinctively stand out in Fig. 7. For convenience, we will use a coordinate (i, j) to refer to the grid in row i and column j. The first distinctive grid is $(12, 23)$, where the Suvarnabhumi International Airport (BKK) is located. This grid is heavily overserved. The mean PDR peaks in the 06:00-11:59 period, then decreases in the afternoon and evening. The second distinctive grid is $(24, 15)$, where the Don Mueang International Airport (DMK) is located. Similarly, this grid is mostly overserved. Like the BKK airport grid, its mean PDR also peaks in the 06:00-11:59 period, then decreases in the afternoon and evening.

V. Conclusion

In this study, we conduct a comprehensive quantitative analysis of how the supply-demand imbalance affects passenger search time and passenger search distance. This has not been investigated in the literature. More than four million Bangkok taxi trips in January, April, July, and October 2022 were thoroughly analyzed. Surprisingly, our analysis shows that the imbalance level and the passenger search time are almost uncorrelated. The correlation coefficient between the imbalance level and the passenger search time is close to zero. Consequently, taxis in underserved (low-demand) areas do not have longer passenger search times than those in overserved (high-demand) areas, and vice versa. This new insight invalidates the conventional perception that taxis would have a harder time finding new passengers in underserved areas. A similar characteristic is also observed in the passenger search distance.

Ultimately, our analysis reveals that, on average, the passenger search times and distances of taxis in underserved and overserved areas are indifferent. While it is more likely to spot passengers in high-demand areas, the taxis have to compete more intensely. As a result, it lengthens the cruising times and distances, making them indistinguishable from those in low-demand areas. These valuable insights provide new

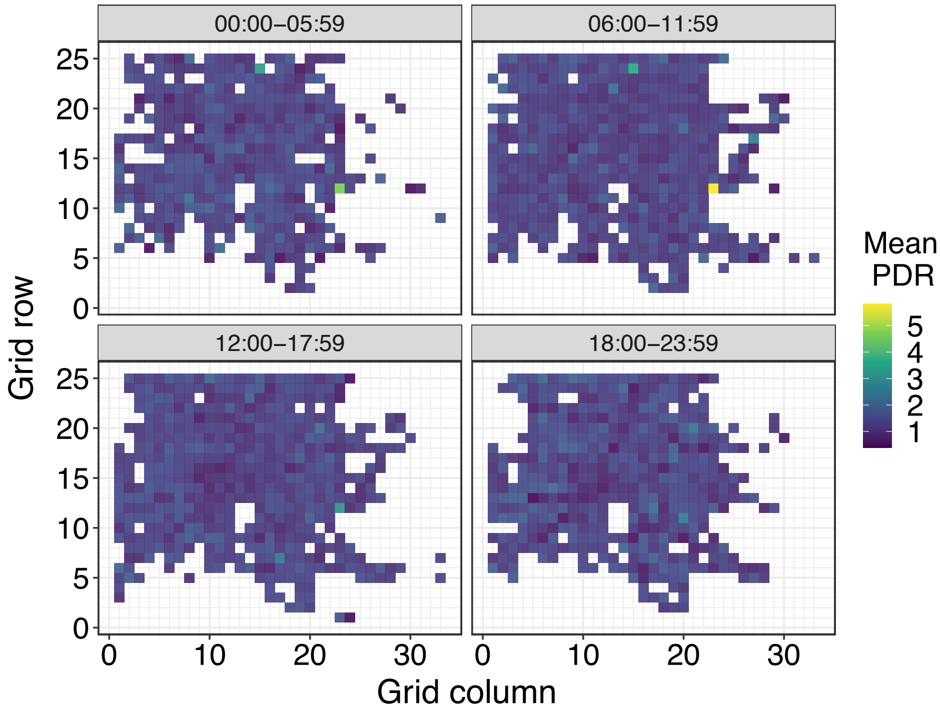

Fig. 7. Mean PDR in each grid cell in four periods of day in July 2022: 1) 00:00-05:59, 2) 06:00-11:59, 3) 12:00-17:79, and 4) 18:00-23:59.

perspectives on designing a recommendation model. It is clear from these insights that recommending taxis to visit a series of attractive high-demand hotspots will not effectively improve the passenger search time and distance. The model should move beyond the simplistic assumption that drivers in underserved areas will have a harder time finding passengers. Instead, the model should consider other factors that practically influence driver behavior, such as profitability, traffic conditions, and driver preferences.

REFERENCES

[1] Intelligent Traffic Information Center Foundation, "Historical raw vehicles and mobile probes data in Thailand," https://iticfoundation.org/service/open-data-sharing/.

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

[3] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Jun. 2013.

[4] A. Saadallah, L. Moreira-Matias, R. Sousa, J. Khiari, E. Jenelius, and J. Gama, "BRIGHT–drift-aware demand predictions for taxi networks," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 234–245, Feb. 2020.

[5] 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.

[6] 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.

[7] P. Shu, Y. Sun, Y. Zhao, and G. Xu, "Spatial-temporal taxi demand prediction using LSTM-CNN," in *Proc. IEEE Int. Conf. Inf. Automat. Sci. and Eng.*, Aug. 2020, pp. 1226–1230.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The rich and the poor: A Markov decision process approach to optimizing taxi driver revenue efficiency," in *Proc. ACM Conf. Inf. and Knowl. Manage.*, Oct. 2016, pp. 2329–2334.

[13] X. Yu, S. Gao, X. Hu, and H. Park, "A Markov decision process approach to vacant taxi routing with e-hailing," *Transp. Res. Part B.*, vol. 121, pp. 114–134, Mar. 2019.

[14] C. Zhang, L. Cheng, Y. Zhang, M. Du, and W. Zheng, "Route optimization of vacant taxicab considering sequential dependence in abstract grid network based on quadtree," *Transp. Lett.*, vol. 16, no. 4, pp. 338–353, 2024.

[15] L. Liu, C. Andris, and C. Ratti, "Uncovering cabdrivers' behavior patterns from their digital traces," *Comput., Environ. and Urban Syst.*, vol. 34, no. 6, pp. 541–548, Nov. 2010.

[16] D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, "Understanding taxi service strategies from taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 123–135, Feb. 2015.

[17] S. Panichpapiboon, "What separates high-performing taxis from the rest? A case study in Bangkok," *IEEE Access*, vol. 12, pp. 194 020–194 029, Dec. 2024.

[18] S. Panichpapiboon and K. Semapakdi-Chang, "Insights from a spatiotemporal analysis of top-performing Bangkok taxis," in *Proc. Int. Conf. Artificial Intell. Inform. and Commun. (ICAIIC)*, Fukuoka, Japan, Feb. 2025, pp. 709–714.

[19] "OpenStreetMap," https://www.openstreetmap.org.

[20] "Open Data Commons," https://opendatacommons.org.