IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 21, 2020, accepted December 29, 2020, date of publication January 8, 2021, date of current version January 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050299

Share-a-Cab: Scalable Clustering Taxi Group Ride Stand From Huge Geolocation Data

WENBO ZHANG[1] AND SATISH V. UKKUSURI[2]
[1]School of Transportation, Southeast University, Nanjing 211189, China
[2]Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47906, USA

Corresponding author: Wenbo Zhang (wenbozhang@seu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 52002064, and in part by the Humanities and Social Science Foundation, Ministry of Education, China, under Grant 20YJC630216.

ABSTRACT Taxi group ride service (TGRS) is one potentially successful way to make traditional services competitive as emerging app-based taxi services, simply through grouping similar taxi rides without significant budget increases, generating one unique pick-up point and one unique drop-off point, thus serving multiple passengers in one single trip. In this study, we mainly develop a scalable method for citywide TGRS stand deployment driven by huge traditional taxicab trips. First, a spatial temporal clustering method is proposed to explore trip clusters that present potential group rides. Second, the agglomerative clustering method is applied to merge trip clusters at both spatial and temporal scale, which will yield potential taxi stand location and schedule. Based on the one-month taxi trips in New York City, the proposed approach can fast process the huge dataset and identify more than 60 stands with four schedules. The study contributes towards efficient methods for developing TGRS in large-scale taxi systems.

INDEX TERMS Data mining, geolocation data, scalable stand deployment, spatial-temporal clustering, taxi group ride services.

I. INTRODUCTION

The last few years have seen a rapid rise in the sharing economy. In transportation, trip sharing and app-based hailing with services such as Uber and Lyft have disrupted traditional transportation services potentially suggesting a future that has important implications for mobility and efficiency of transportation services in urban areas. One of potential implications is to group rides in a way that is beneficial both for the users (reduced cost) and the system (reduced congestion). These emerging services that are always convenient and competitive have challenged the traditional street-hail taxi service. Confronted with challenges, city policymakers and transportation authorities are devoting remarkable efforts to make these services competitive for users. Taxi group ride service (TGRS) is one potentially successful way by grouping similar taxi rides without significant budget increases. Under TGRS, taxicabs can pick up more than one passenger at a fixed taxi stand and drive to a predefined drop-off area, likely a street or a district. Passengers will pay with the fixed fare rate per trip, not a distance-based and travel time-based fare rate. In addition, the passenger group should be self-organized only at the fixed taxi stand, likely with friends or strangers, and no more pickups are allowed along the route. Figure 1 shows one taxi group ride stand in New York City (NYC) and its corresponding regulations. The TGRS is different from the dynamic taxi ride sharing defined in [1] that can pick up anywhere along one route.

In this study, we mainly focus on taxi stand deployment of TGRS while planning. The basic problem discussed is to identify frequent taxi group rides and locate corresponding stands in a large-scale taxi system with the mixture of traditional street-hail taxi service and TGRS. The study is developed based on the following key observations:

- Compared with dynamic taxi ride sharing, TGRS is easily regulated and implemented. Street-hail taxi service mixed with TGRS can also meet various demand for taxi services and perform at comparative level of service even without dispatching/recommender system. Moreover, gaps in techniques (e.g. communication, information, and computation) and substantial budget impede explosive growth in dynamic taxi ride sharing. Thus, TGRS may be a better option for improving street-hail taxi service;

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

FIGURE 1. A typical taxi stand and corresponding regulations in NYC.

- In big cities, it is more likely to have frequent taxi flows within downtown areas especially during peak hours (see Figure 2). This is an important concern for developing TGRS for large-scale taxi systems. Taxi systems may serve more passengers during peak hours, save passengers' trip cost, improve drivers' revenue, reduce traffic flows on corresponding routes, and possibly decrease congestion and emissions by grouping similar taxi trips. Both NYC and Beijing, China proposed initiatives on TGRS and implemented testing programs around few hotspots. The positive effects of TGRS may encourage cities to develop large-scale TGRS programs; and
- The availability of pervasive data, such as GPS-based taxi trip information, provides a new perspective on exploring distribution of taxi demand and supply, mining frequent taxi trips, and measuring the potential of TGRS in a large-scale taxi system. Various characteristics of taxi movements have been obtained based on similar datasets, which are impossible to apply for traditional datasets [2]–[10].

In recent years, given the interest in improving the efficiency of taxi services, ride sharing is a topic of growing interest. However, most studies focused on dynamic taxi ride sharing, not on TGRS [1], [7]. Furthermore, three limitations in the studies limit the value of past studies. First, previous work does not fully consider the slugging form of ridesharing (i.e. passengers walk a distance to nearest pickup location, take the group ride, and walk to the destinations from drop off location). Zhan *et al.* [7] made a preliminary discussion on ridesharing and proposed k-matching model to simulate ridesharing behaviors. However, the ridesharing system is dynamic one which is different from the slugging one. Ma and Wolfson [4] focused on the slugging form of ridesharing, but mainly addressed the efficiency of this form. Second, the computation while addressing the large-scale problem is demanding, regardless of the method in [1] and graph-based approach in [7].

The study contributes towards a fast approach for large-scale TGRS stand deployment from traditional street-hailing taxi trips. To process the spatial and temporal characteristics of taxi trips in this problem, a two-stage modeling structure is proposed including a spatiotemporal clustering to identify trip clusters and an agglomerative clustering method for merge trip clusters. One-month (September, 2015) taxi trips in NYC is introduced to validate proposed modeling structure. The following sections are organized as follows: section II presents the modeling structure, definitions, and equations; section III shows the case study and results; section IV concludes this study and points out future study.

II. METHODS

A. MODELING STRUCTURE

The street-hailing taxi trip x_n is a time-stamped record of 5-tuples $\left(O_{lat}, O_{long}, D_{lat}, D_{long}, T\right)_n$, where O_{lat}, O_{long} represent the latitude and longitude of origin, D_{lat}, D_{long} represent the latitude and longitude of destination of each taxi trip, and T represents departure time. The problem is to deploy the TGRS stands with enough taxi rides in most days. Hence, a modeling structure with two stages (see Figure 3) are proposed to process spatial and temporal clustering and large-scale issue:

- Cluster street-hail taxi trips with similar origins, destinations, and departure time that may match together for a group ride. In this step, we view each existing street-hail trip as a point with both spatial and temporal attributes (i.e. origin, destination, and departure time). A spatial-temporal density-based spatial clustering of applications with noise (ST-DBSCAN) is proposed to cluster points; and
- Identify placement of TGRS stands by merging street-hail taxi trip clusters. The agglomerative clustering method is applied for trip clusters from ST-DBSCAN at spatial scale to identify the spatial location of TGRS stands. Then the same method is applied at the temporal scale to identify the schedule of TGRS stands (i.e. all day, peak hours, morning peak, and evening peak).

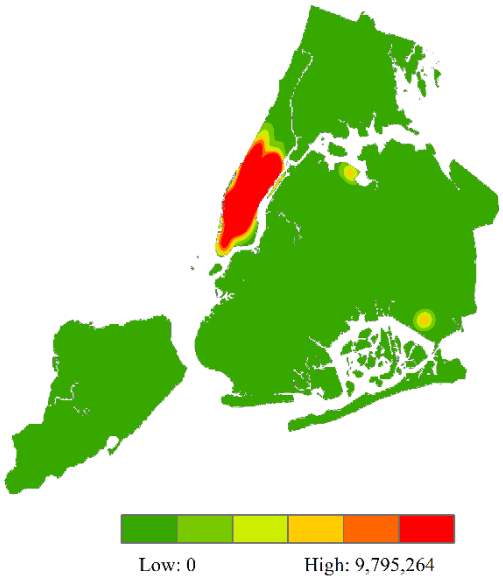
B. ST-DBSCAN

1) REVISIT ON DBSCAN

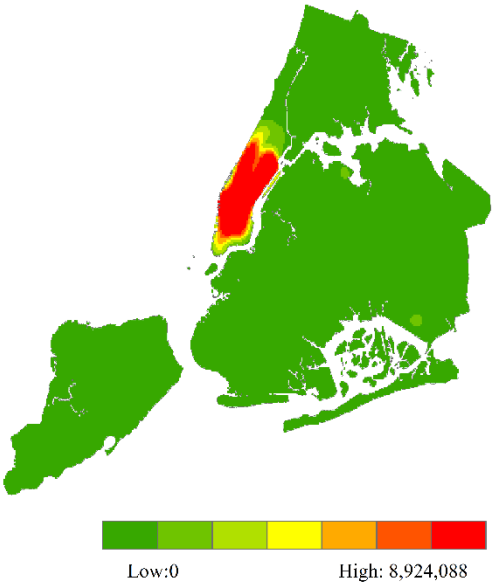
Let $\mathrm{X} = \{x_1, x_2, \cdots, x_n\}$ be the object set and $d(\cdot, \cdot)$ is a metric distance.

Definition 1 (Core Object): The object x_p is a core object if its $\varepsilon-$ neighborhood contains at least *num* many objects. That is, $N_\varepsilon\left(x_p\right) = \left\{y \in X | d\left(y, x_p\right) \leq \varepsilon\right\}$ and $\left|N_\varepsilon\left(x_p\right)\right| \geq num$.

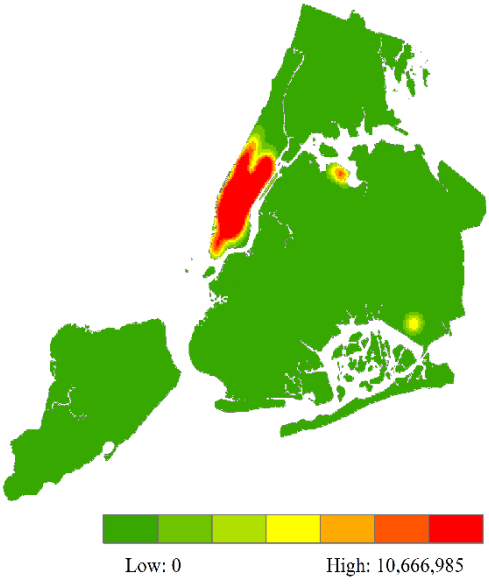
Definition 2 ($\varepsilon-$Reachable): Two core objects are $\varepsilon-$ reachable if each core object is in $\varepsilon-$ neighborhood of another core object. That is, $x_p \in N_\varepsilon(x_q)$ and $x_q \in N_\varepsilon(x_p)$.



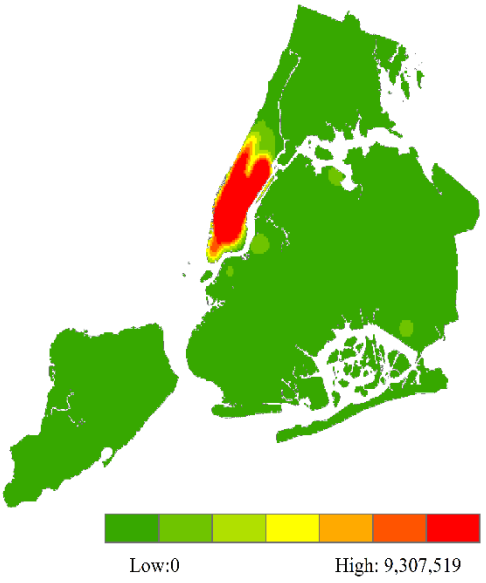
(a) taxi trip origin density on 9/10/2015



(b) taxi trip destination density on 9/10/2015



(c) taxi trip origin density on 9/20/2015



(d) taxi trip destination density on 9/20/2015

FIGURE 2. **The trip density between 18:00 and 19:00 in NYC.**

Definition 3 (Density Connected): two core objects are density connected if they are directly or transitively $\varepsilon-$ reachable.

Definition 4 (Cluster): A cluster C is a non-empty maximal subset of X such that every pair of objects in C is density connected.

2) ST-DBSCAN

The DBSCAN cannot process both spatial and temporal dissimilarity (i.e. metric distance) simultaneously. To improve this weakness, we introduce one more ε for temporal dissimilarity [11]. Thus, the $\varepsilon-$ neighborhood of object x_p is:

$$N_\varepsilon\left(x_p\right) = \left\{y \in X | d_{spatial}\left(y, x_p\right) \leq \varepsilon_{spatial}\right\} \cap \left\{y \in X | d_{temporal}\left(y, x_p\right) \leq \varepsilon_{temporal}\right\} \quad (1)$$

The spatial distance can be estimated by the Manhattan distance. The temporal distance is the difference between departure times. To accelerate computation of metric distance, the spatial distance is derived by Manhattan distance if one object x_q is in a certain range of another object x_p. The object

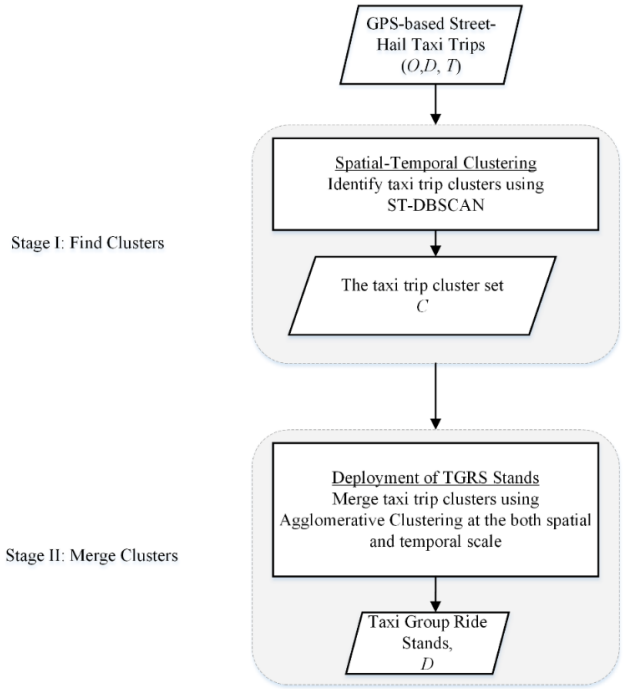


FIGURE 3. The two-stage approach for TGRS stand deployment.

x_p is the centroid of a cube with origin difference $2km$, destination difference $2km$, and departure time difference 10 min. Or the spatial distance is set as a big value M.

C. AGGLOMERATIVE CLUSTERING

To improve computational load, the spatial centroid of each trip cluster is utilized to represent all trips in the cluster and the range of departure time is introduced to replace the departure time of each trip in the cluster.

The next step is to merge the trip clusters by the complete linkage method at the spatial scale, as well as temporal scale. Two criterions are applied:

- Suppose each taxi stand can serve an area with radius of $1km$. Thus, we can merge the trip cluster based on spatial distance till the spatial distance of any pair centroid in the merged cluster is greater than $2km$.
- The merged cluster should be presented at least 70% of total days during test period. Or the merged cluster may fail to provide enough group rides in most days. In addition, this check will avoid temporary many trips induced by events.

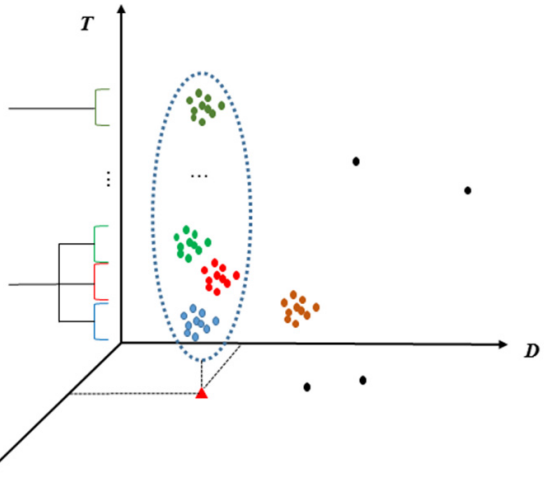


FIGURE 4. The illustration of cluster merge for taxi stand (each color represents one subgroup of similar trips).

FIGURE 5. The centroids of trip clusters from ST-DBSCAN.

See Figure 4, the aforementioned merging process is illustrated in the dashed circle. The trip clusters in the circle are with similar spatial locations, thus we can derive a taxi stand (i.e. the red triangular) by finding centroid of the merged cluster.

$$d_{spatial}\left(x_p, x_q\right) = \begin{cases} Manhattan\left(O(x_p), O(x_q)\right) + Manhattan\left(D(x_p), D(x_q)\right) & if\ x_q \in R\left(x_p\right) \\ M & \text{otherwise} \end{cases} \tag{2}$$

$$R\left(x_p\right) = \left\{ (O, D, T) \left| \begin{array}{c} O_{lat} \in \left[O_{lat}(x_p) - 0.008, O_{lat}(x_p) + 0.008\right] \\ O_{long} \in \left[O_{long}(x_p) - 0.005, O_{long}(x_p) + 0.005\right] \\ D_{lat} \in \left[D_{lat}(x_p) - 0.008, D_{lat}(x_p) + 0.008\right] \\ D_{long} \in \left[D_{long}(x_p) - 0.005, D_{long}(x_p) + 0.005\right] \\ T \in \left[T(x_p) - 5, T(x_p) + 5\right] \end{array} \right. \right\} \tag{3}$$

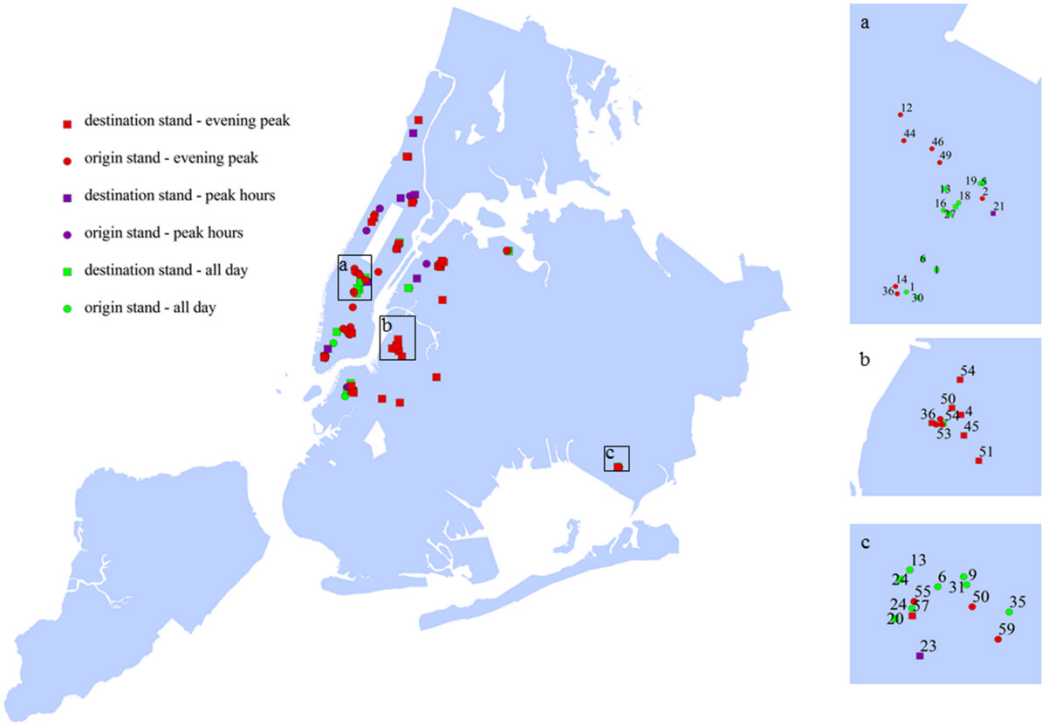


FIGURE 6. The TGRS taxi stands.

Next step is to process the temporal interval of trip clusters and identify the schedule of each taxi stand. If two intervals overlap, we can derive the union of two intervals as one new interval (see the dendrogram on the time axis in Figure 4). Till there are no overlaps of temporal intervals in each merged cluster, we can summarize the percentage of covered times in one day, the percentage of covered peak hours (including morning peak 7am to 9am and evening peak 5pm to 7pm), the percentage of covered morning peak hours, and the percentage of covered evening peak hours. The four types of schedule (i.e. all day, peak hours, morning peak, and evening peak) are determined based on the following criterions:

- If final temporal interval covers more than 50% of times of one day, the TGRS stand can be operated in the whole day; and
- If fail to operate in the whole day, we can check the peak hours. If the temporal interval can cover more than 50% of both peak hours or AM/PM peak hours, the TGRS stand can operate during both peak hours, AM peak, or PM peak, respectively.

III. RESULTS

We have collected 2015 GPS-based street-hail taxi trip dataset from NYC taxi and limousine commission. However, only one-month trips in September, 2015 will be extracted as a sample, considering the enormous taxi ridership of 1.8 billion and comparative monthly distribution of taxi demand in the large-scale taxi system. Unlike free moving objects, the taxi movements are constrained by urban road network configuration. A general approach to match all GPS records onto road network should be introduced to measure exact spatial relationship between two points. However, the map-matching process is computationally expensive with enormous locations in a large-scale taxi system. In the case study, we estimate distance based on Manhattan distance that is a good approximation of the real distance, instead of map-matching.

TABLE 1. Summary statistics of all TGRS stands.

Type of TGRS stand	No.
All-day	22
Peak hours	8
AM peak	0
PM peak	31
total	61

In the stage of ST-DBSCAN, we set $\varepsilon_{spatial} = 2km$, $\varepsilon_{temporal} = 5min$, and $num = 10$. For 10,718,718 street-hail taxi trips, the ST-DBSCAN identifies 8,751 trip clusters covering 9,333,934 street-hail taxi trips. Only 13% of street-hail taxi trips are measured as noise. Continuing to the second stage of merging trip clusters, we can derive 271 merged cluster, but only 61 out of which are with consecutive group rides in more than 20 days (i.e. 70% of all days during test period). Based on summary on merged temporal intervals, we can determine the type of each TGRS stand, shown in Table 1. The distribution of all 61 stands and corresponding types are

shown in Figure 6. The stand deployments in three typical regions (i.e. Manhattan midtown, Brooklyn downtown, and JFK airport) are in details in subplots of Figure 6.

IV. CONCLUSION

This study mainly focuses on TGRS in a large-scale system. Considering spatiotemporal characteristics and big data, we develop a fast two-stage approach for TGRS stand deployment. This approach can reduce computational load by firstly finding trip clusters and corresponding centroids. Then the agglomerative clustering can efficiently identify the taxi stand and corresponding schedule by merging trip clusters at both spatial and temporal scale. The case study in NYC confirms the performance and feasibility of proposed modeling structure. Furthermore, this study can be improved in two ways: a) the sensitivity analysis on predefined input parameters; and b) consider the number of street-hail taxi trips in each trip cluster while finding the centroids of merged clusters.

REFERENCES

[1] S. Ma, Y. Zheng, and O. Wolfson, "T-share: A large-scale dynamic taxi ridesharing service," in *Proc. IEEE 29th Int. Conf. Data Eng. (ICDE)*, Brisbane, QLD, Australia, Apr. 2013, pp. 410–421.
[2] F. He and Z.-J.-M. Shen, "Modeling taxi services with smartphone-based E-hailing applications," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 93–106, Sep. 2015.
[3] C. Kamga, M. A. Yazici, and A. Singhal, "Analysis of taxi demand and supply in new york city: Implications of recent taxi regulations," *Transp. Planning Technol.*, vol. 38, no. 6, pp. 601–625, Aug. 2015.
[4] S. Ma and O. Wolfson, "Analysis and evaluation of the slugging form of ridesharing," in *Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Orlando, FL, USA, Nov. 2013, pp. 64–73.
[5] L. M. Martinez, G. H. A. Correia, and J. M. Viegas, "An agent-based simulation model to assess the impacts of introducing a shared-taxi system: An application to lisbon (Portugal)," *J. Adv. Transp.*, vol. 49, no. 3, pp. 475–495, Apr. 2015.
[6] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti, "Quantifying the benefits of vehicle pooling with shareability networks," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 37, pp. 13290–13294, Sep. 2014.
[7] X. Zhan, X. Qian, and S. V. Ukkusuri, "Graph-based approach to measuring efficiency of urban taxi service system," in *Proc. 94th Annu. Meeting Transp. Res. Board*, Washington, DC, USA, 2015, pp. 1–5.
[8] W. Zhang, X. Qian, and S. V. Ukkusuri, "Identifying the temporal characteristics of intra-city movement using taxi geo-location data," in *Enriching Urban Spaces with Ambient Computing, the Internet of Things, and Smart City Design*, S. Konomi and G. Roussos, Eds. Hershey, PA, USA: IGI Global, 2016.
[9] W. Zhang and S. V. Ukkusuri, "Optimal fleet size and fare setting in emerging taxi markets with stochastic demand," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 31, no. 9, pp. 647–660, Sep. 2016.
[10] W. Zhang, S. V. Ukkusuri, and J. J. Lu, "Identifying the determinants of the empty taxi trip duration using limited geo-location data," *Transportation*, vol. 44, pp. 1445–1473, Dec. 2016.
[11] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, Jan. 2007.

WENBO ZHANG received the B.S. degree in transportation from the China University of Mining and Technology, Xuzhou, China, the M.S. degree in transportation engineering from Southeast University, Nanjing, China, and the Ph.D. degree in civil engineering from Purdue University. He is currently an Assistant Professor with the School of Transportation, Southeast University. His research interests include transportation big data analytics, urban mobility analysis, taxi system modeling and optimal control, and complex networks.

SATISH V. UKKUSURI is currently a Professor with the Lyles School of Civil Engineering, Purdue University, and he is also recognized nationally and internationally in the areas of transportation network modeling and disaster management. He is the Director of the Interdisciplinary Transportation Modeling and Analytics Lab, Purdue University. He has published extensively on these topics in peer reviewed journals and conferences. His current interests include dynamic network modeling, large-scale data analytics, disaster management issues, and freight transportation and logistics.