IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 25, 2019, accepted May 30, 2019. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.2922210

Constructing Time-Dependent Origin-Destination Matrices With Adaptive Zoning Scheme and Measuring Their Similarities With Taxi Trajectory Data

WERABHAT MUNGTHANYA[1], SANTI PHITHAKKITNUKOON[1], MERKEBE GETACHEW DEMISSIE[2], LINA KATTAN[2], MARCO VELOSO[3], CARLOS BENTO[3], AND CARLO RATTI[4]

[1]Department of Computer Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
[2]Department of Civil Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
[3]Center for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
[4]SENSEable City Laboratory, Massachusetts Institute of Technology, MA 02139, USA

Corresponding author: S. Phithakkitnukoon (santi@eng.cmu.ac.th)

This research was supported by the Eyes High Postdoctoral Fellowship Program, Alberta Transportation, and Alberta Motor Association (AMA).

ABSTRACT There has been a recent push towards using opportunistic sensing data collected from sources like automatic vehicle location (AVL) systems, mobile phone networks, and global positioning system (GPS) tracking to construct origin-destination (O-D) matrices, which are an effective alternative to expensive and time-consuming traditional travel surveys. These data have numerous drawbacks: they may have inadequate detail about the journey, may lack spatial and temporal granularity, or may be limited due to privacy regulations. Taxi trajectory data is an opportunistic sensing data type that can be effectively used for O-D matrix construction because it addresses the issues that plague other data sources. This paper presents a new approach for using taxi trajectory data to construct a taxi O-D matrix that is dynamic in both space and time. The model's origin and destination zone sizes and locations are not fixed, allowing the dimensions to vary from one matrix to another. Comparisons between these spatiotemporal-varying O-D matrices cannot be made using a traditional method like matrix subtraction. Therefore, this paper introduces a new measure of similarity. Our proposed approaches are applied to the taxi trajectory data collected from Lisbon, Portugal as a case study. The results reveal the periods in which taxi travel demand is the highest and lowest, as well as the periods in which the highest and lowest regular taxi travel demand patterns take shape. This information about taxi travel demand patterns is essential for informed taxi service operations management.

INDEX TERMS Dynamic origin-destination matrix, adaptive zoning scheme, origin-destination matrix similarity measure, taxi trajectory data, taxi travel demand.

I. INTRODUCTION

A. BACKGROUND

The demand for transport is derived by the locations people travel for different activities such as work, leisure, health, etc. To understand the demand for transport, we must understand the spatial and temporal distributions of trips and the locations of the activities they serve. This information is summarized in Origin-Destination (O-D) matrices, which contain information about the spatial and temporal distributions of activities between different traffic zones (TAZs) in an urban area. Each cell in the matrix represents the number of trips departing at a given time interval and transiting between an origin and a destination within the study area.

O-D estimation has previously been used to provide the necessary input for long-term strategic planning, as well as for short-term transportation planning purposes. Hellinga [1] suggested four different ways of categorizing existing O-D estimation approaches to provide a structure within which an assessment of different O-D estimation approaches can

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

be made. Distinctions are made between heuristic and mathematical approaches, O-D estimations for specific area versus for general networks, static versus dynamic O-D estimations, and the approaches' assumptions regarding the types of routes utilized by drivers.

This paper presents a new approach for using taxi trajectory data to construct a taxi O-D matrix that is dynamic in both space and time. The distinction between static and dynamic O-D estimation approaches depends on the consideration of temporal variations [2]. A static O-D matrix represents the number of trips between origins and destinations over a relatively large period within which steady-state traffic conditions are assumed [3]. This is not a good approximation of reality, where there exists a definite temporal variation of O-D flows over the course of a large analysis period. Thus, for any short-term planning study, a knowledge of time-dependent O-D flows could be very practical [2]–[4].

Mobility need manifests through a trip between origin and destination locations. A disaggregate analysis of trips to model the transportation system would result in huge difficulties [5]. Practically all transport models make use of TAZs for aggregate computations on groups of locations and individuals. The scientific literature specifies that transport demand modeling requires consistent zonal data aggregation along the entire stage of travel demand forecasting modeling [6]. This fixed-zone based system is especially important for developing and maintaining regional travel demand forecasting models, which analyze and forecast the volume of all modes of travel in a region for both people and freight. In many cases, estimation of O-D demand is necessary for the entire study region. In other cases, O-D demand is estimated only for a portion of the regional or metropolitan area network rather than for the entire network [4].

The fixed-zonal level travel demand modeling approach is not without its limitations. For example, a trip distribution model that assumes that trip origins and destinations are concentrated around the zone's centroid ignores intrazonal flows, whose trip distances are always positive, because the separation for these flows would be zero [7], [8]. Using the zone's centroid as an approximation creates bias during traffic assignment because it overestimates local traffic near the zone centroid and underestimates it elsewhere [9]. The computational requirements of spatial interaction models typically rise with the square of the number of zones [10]. The issue of how to deal with a very large number of spatial choice alternatives for destination choice models is another longstanding problem [11].

The adaptive zoning method has recently been explored to address these problems. In contrast to traditional zoning, where study regions are partitioned into predefined and fixed zones for the development stages of the entire model, adaptive zoning establishes a collection of different zone plans. This approach has improved the scalability of spatial interaction models [10], road traffic assignments [9], and mode choice modeling [12]. Ben-Akiva and Lerman [13] suggested using a restricted set of zonal alternatives rather than a full set when developing a destination choice model. Hammadou *et al.* [11] also suggested using a reduced number of zones that are defined based on criteria, suggesting areas of homogeneous land use.

Estimation and prediction of time-dependent O-D flows have gained further relevance in the context of Intelligent Transportation Systems (ITS). Some of the essential features of an ITS are facilitating real-time routing policies, traffic management and control, and traffic information provisions capable of achieving system-wide objectives [1], [3], [4], [14]. One of the most important components of an ITS is the time-dependent O-D estimation and prediction module [14]–[19].

Data for time-dependent O-D estimation and prediction come from various sources. The most commonly-used procedures for obtaining time-dependent O-D flows are surveys (e.g., household, roadside). These procedures are expensive, time consuming, and can be difficult to carry out. In the absence of direct observation, O-D flows are estimated based on zonal land use data such as population and employment opportunity [20]. Another approach is to infer unknown O-D demand from observed link traffic flow data, where measurement can be made using loop detectors, video cameras, etc. [2], [21], [22]. In a dynamic or time-dependent O-D context, the estimation procedure would also include prior O-D information, which typically comes from results of previous estimation or historical database of time-dependent O-D flows [1], [3], [4], [14]–[16].

B. MOTIVATION FOR THE NEW APPROACH

Due to the variation of activity density in urban areas, the demand for different transport services shows periodicity (e.g., weekly, daily, hourly) [23], [24]. Like with other transportation systems, demand for taxi services exhibits periodicity in time and space that reflects the underlying patterns of human activity [25]. The previous gap in our understanding of the temporal and spatial variations of taxi demand was primarily a result of data limitations [6]. The taxi industry has recently collected a large volume of taxi trajectory data that can be used to infer the origins and destinations of taxi trips.

One method of understanding the origins and destinations of taxi trips is pick-up and drop-off hotspot detection at major taxi trip generation and attraction areas [6], [26]–[30]. To improve taxi services, it is important to understand taxi demand, how that demand varies through space and time, and which attributes influence that demand. Taxi trip generation models provide an idea of the level of taxi trip attraction and generation rates in a certain area. Lacombe *et al.* [31] and Yang *et al.* [32] developed two taxi trip generation models, one for trip production and the other for trip attraction, to achieve a better understanding of taxi demand. They applied various explanatory variables such as demographics, land use, accessibility to transit, and weather conditions to those models to determine whether any of those factors were likely to influence taxi demand. Zhang *et al.* [33] and

Liu *et al.* [34] extended these models by extracting the O-D trips between trip generation and trip attraction areas. Further improvements on the aforementioned models were achieved through the development of models that distribute taxi trips obtained from a trip generation model among destinations [35], [36]. These studies consider static (time-independent) taxi O-D flows and thus do not capture the dynamics of time-varying taxi O-D flows or of changes in the spatial distribution of taxi O-D flows over time. There is a clear need for increased knowledge about taxi O-D estimation, especially in characterizing taxi O-D flows for different times of the day.

Defining taxi pick-up and drop-off zones, or origin and destination zones, respectively, is a challenging task when developing a taxi O-D estimation model. In the case of Lisbon city, estimating taxi O-D at a census block level is difficult because of the high number of alternatives (there are 3,712 census blocks in Lisbon). The choice of administrative districts in Lisbon such as freguesias (parishes) results in large sized zones, and thus, the number of intra-zonal trips is substantial. This is especially important when a high share of taxi trips is short and could result in a significant number of intra-zonal trips [37].

Academic consensus is that spatial structures of significant trip generators in a fixed zonal level system are long lasting, masking the variability and evolution of the activities people perform at these locations [38], [39]. Urban areas continually change, so the same urban structure may generate different levels of trips over time. One example of this type of change is a transit station that sees an increase in traffic due to an event, a trendy new cafeteria or bar, a recently-renovated building that attracts new companies, etc. [40]. Sevtsuk and Ratti [41] argue that there are sequences of urban activities that take place at varying times of day that affect the dynamics and forms of urban areas. In the context of taxi services in Lisbon, for example, office areas are known to have a high number of taxi passenger drop-off and pick-up events in the daytime and areas of the city with bars and nightclubs exhibit a reverse pattern. Another example is the Lisbon international airport, where taxi travel demand tends to follow the peaking characteristics of passenger enplanements and passenger deplanements and tends to be more evenly distributed during off-peak hours [37].

Typically, a zoning plan for a city's fixed zonal level scheme is defined such that the downtown area (or the area of a city with a high flow density) is represented in more detail. Zones become progressively larger moving away from the downtown [10], [42], [43]. In this case, the zoning scheme does not account for time-sensitive changes to the trip generation and attraction roles of each zone. The zoning scheme must be improved to adapt the sizes of the origin and destination zones to the volume of taxi trips moving in and out of each zone based on the time of day.

We attempt to address these challenges through a new framework comprised of time-dependent taxi O-D matrices with adaptive zoning schemes that reflect the continuously-changing demand for taxi travel. The zoning scheme is defined such that urban areas with the strongest taxi movements are represented in more detail and the degree of aggregation becomes higher when the demand for taxis is low. We also develop a similarity measure framework to compare matrices of different dimensions. In our case, the framework addresses the different number of zones between O-D matrices due to the implementation of adaptive zoning because conventional similarity measurements like matrix subtraction, correlation coefficient (R-squared), Geoffrey E. Havers (GEH) statistic [44], Root Mean Square Error (RMSE), and Eigenvalue-based measure (EBM) [45] cannot be directly applied due to their common requirement of dimension equality for all matrices.

Another possible approach is utilizing the matrix norms [46]. Matrix norm is a vector norm in a vector space whose elements are matrices. Vector norm is a function that assigns a positive length or size to each vector in a vector space. A single positive value or matrix norm would be calculated and representing a given matrix based on which a similarity across different matrices could be measured. However, a lot of information (e.g., flow, direction, so on) would be lost through the compression of matrix elements into a single value, i.e., a matrix norm. To the best of our knowledge, no study has attempted to measure similarities between O-D matrices of different dimensions.

This paper makes three important contributions. First, we develop an adaptive zoning scheme that could potentially support transportation operations decisions for situations that may not require analysis through a traditional fixed zoning system or a complete network representation. Second, we create time-dependent O-D matrices with adaptive zoning that cannot be compared using a traditional approach like matrix subtraction. We therefore develop a method to measure similarities across these doubly-dynamic O-D matrices. Third, we analyze the self-similarities and cross-similarities of the doubly-dynamic O-D matrices. Our objective is to draw actionable insights from taxi travel demand across different time periods. Taxi operators and transport planners can use these insights to establish better operational strategies to improve taxi services.

II. RELATED WORK

Recently, focus has shifted towards using opportunistic sensing datasets produced from various sources that can provide insights into the spatial distribution and temporal evolution of the movements of people and vehicles in cities. Opportunistic sensing data is collected for one purpose, but also creates an opportunity for another purpose. For example, the Automatic Vehicle Location (AVL) system has been widely deployed in public transportation systems so that the locations of public transit vehicles are known. This data is collected via GPS and a transmission mechanism to facilitate dispatching and fleet management. Combined with Automated Fare Collection (AFC), AVL data gives movement information about people, and is therefore opportunistically a source of data that can also be used to describe citywide mobility [47]–[50].

Although AVL and AFC data from public transportation such as buses and trains can yield useful data and information on general passenger movement, it does not provide the exact origin and destination for each passenger since these transportation modes operate on pre-designated stops and routes on fixed schedules.

Another type of opportunistic sensing data is mobile phone call detail records (CDRs), which are users' communications and corresponding location records. When a mobile phone user connects to the cellular network by making or receiving a phone call or using the internet, the communication (e.g., call duration, timestamp, caller's and recipient's identifications) and location of the connected cellular tower are recorded for billing purposes. The CDR location records of individual users have been used in human mobility studies. The use of cellular network data has been explored for the development of large-scale mobility sensing since the early 2000s [51]. The data has been used to investigate various aspects of transportation issues, including large-scale urban sensing [41], [52], [53], traffic parameter estimation [54]–[56], commuting trip estimation [57], [58], transport mode choice [59], and land use inference [60], [61]. In spite of the CDR data's versatility in support of a wide range of studies, the data come with two main limitations [62]: CDR is sparse, as it is only acquired when a device connects to the cellular network; and CDR data is spatially coarse because the location record is only available at the granularity of cell tower service coverage. The CDR data can be used to infer about commuting trips convincingly [57], [58] but not for non-commuting trips e.g., leisure and touristic trips.

Tracking an individual with a GPS-enabled device can provide more detailed information about that person's movement than CDR data. Collectively, this approach can comprehensively address almost all data requirements of each of the step in the four-step travel demand model, except for information about the transportation mode. Due to privacy issues and regulations like the EU general data protection regulation, however, collecting such data on a large scale is difficult and challenging. Recent attempts have produced datasets that are limited to the specific type of tracked individuals, such as university students [63] or customers of a particular service provider where the data was obtained in exchange for incentives [64]. Privacy concerns largely prevent this type of detailed mobility data from being publicly available or from being extensively utilized, meaning that it is not easily exploitable for O-D matrix estimation.

Aggregate trip datasets, including O-D flow data, have recently been made available from companies such as Google, INRIX, HERE, etc. These datasets offer many advantages, including road link O-D flow, travel time, congestion index, etc. They allow transportation and planning agencies to track trends and calibrate models for more informed decision making. However, the aggregate nature of the data does not support a fine-grained analysis of trips by different modes. For such an analysis, researchers must turn to datasets that collect information at the individual level.

Dynamic O-D flows are an essential component for the success of intelligent transportation systems and advanced traveler information systems [14]–[16]. Vehicle-based detection techniques are recent data sources for estimating dynamic O-D flows. These can include beacon-based probe vehicles [17]–[19], active probe vehicles such as floating car data [65]–[67], and passive probe-vehicles [68], [69]. Data from probe vehicles does not provide a complete record of traffic volume because of only a fraction of the traffic has been equipped [65] and because fleet probe vehicles (e.g., transit and taxi vehicles) may not be representative of the global mobility pattern since these vehicles operate for other primary purposes and could therefore inherit highly-biased traffic characteristics [70], [71].

A taxi is a passive probe vehicle that represents a type of public transportation that is different from fixed schedule and route services since taxi pick-up and drop-off locations are determined by passengers. AVL data from a taxi can provide detailed information regarding the origins and destinations of the cab's trips. The process of data collection is transparent and non-intrusive to passengers since no personally-identifying information is recorded. Only the taxi's GPS location and occupancy status are recorded for dispatching and management purposes; hence, there is no concern for privacy issues. In addition to the extraction of trip pick-up (origin) and drop-off (destination) information, taxi trajectory data has also been used as a probe to monitor road traffic conditions [72]–[75] and to understand urban dynamics [64], [76], [77]. It enables us to take a collective snapshot of urban movement, giving us an overview of how the city functions economically and socially [78], [79].

This paper attempts to create a better understanding of taxi travel demand through the development of a framework to construct a dynamic O-D matrix that is a spatiotemporal-variant description of ever-changing taxi travel demand. This O-D matrix is unlike existing dynamic O-D matrices that are time-variant on fixed zonal areas [80]–[82]. Since our dynamic O-D matrix is time- and space-dependent, creating output matrices of different dimensions, comparisons between different matrices cannot be made directly with a traditional approach like matrix subtraction. We have addressed this by developing a method to measure similarities across dynamic O-D matrices that are time-dependent with adaptive zoning schemes.

III. DATASET

We used a set of taxi trajectory data collected by Geotaxi, one of the main taxi service providers in Lisbon, Portugal. The dataset includes data from 172 taxicabs over a two-month period (September and October 2009), amounting to nearly three million taxi location traces. The data sampling rate varies according to the trip – i.e., distance driven, time elapsed, and service status changed (occupied, vacant).

Lisbon is the capital of Portugal. Lisbon's urban area, which expands around the downtown core, has a high population density and boasts touristic, historic and commercial areas, and public transportation hubs. Residential areas and airport and industrial facilities are located in the city's periphery. In 2009, Lisbon had a population of 484,723. According to a report by Darbéra [83], taxis were used for a majority of trips made by tourist visitors. The top purposes for taxi usage in Lisbon were leisure, work, business, medical care, and airport.

The data from each of the 172 taxis carries information about the taxi's location and service status, and a corresponding date and time. If $S = \{s_1, s_2, \ldots\}$S $= \{s_1 s_2, \ldots\}$ represents a trace of a taxi, then each instance sample k contains a location, service status, and timestamp: $s_k = (latitude_k, longitude_k, service_k, timestamp_k)$.

FIGURE 1. **A sample taxi trace over a five-hour period in service. Red dots represent recorded locations while the taxi is occupied and green dots indicate that the taxi is vacant.**

Figure 1 shows a sample trace of a single taxi over its five hours in service. Each dot represents the recorded location in either red or green to denote a service status of occupied or vacant, respectively.

FIGURE 2. **Methodology flowchart.**

IV. METHODOLOGY

Our methodology included the construction of a time-dependent O-D matrix based on an adaptive zoning scheme. The adaptive zoning scheme yielded matrices of different sizes, so we needed to devise a method of measuring similarity between the output O-D matrices. Figure 2 shows the overall process flow, which includes gathering taxi data, detecting pick-up and drop-off locations based on the service status information, clustering the origins and destinations using the X-means algorithm, filtering out potential outliers using DBSCAN, identifying areas of origin and destination concentrations using convex hull, time-dependent O-D matrix construction, and measuring similarities across the time-dependent O-D matrices with adaptive zoning schemes.

A. O-D MATRIX CONTRUCTION

The key data elements in the development of travel demand modeling are trips between origins and destinations. The construction of a basic O-D matrix is the starting point of the O-D matrix estimation. We use taxi GPS data to construct an initial O-D matrix that reflects taxi trip patterns and that provides information related to the origins, destinations, and volumes of taxi trips. A priori O-D matrices are traditionally obtained from historical O-D tables (previous surveys), which tends to be costly, labour intensive, and time disruptive to the trip makers. In our analysis, trip makers' pick-ups and drop-offs are used to locate the trip's origins and destinations.

A taxi meter's status transitioning from one state to the other is used to determine passenger pick-up and drop-off events as well as origin and destination locations. A pick-up location can be identified when the service status changes from available to occupied as the taxi picks up a new passenger. A drop-off location can similarly be identified when the service status switches from occupied to available as the passenger is dropped off. A set of pick-up locations is a subset of S and can be denoted as $P = \{s_k \in S | service_k = \text{“}occupied\text{”}\ and\ service_{k-1} = \text{“}available\text{”}\}$. Likewise, a set of drop-off locations (D) is a subset of S and can be denoted as $D = \{s_k \in S | service_k = \text{“}available\text{”}\ and\ service_{k-1} = \text{“}occupied\text{”}\}$. This intuitive approach allowed us to identify a total of 101,463 trips, each comprised of a pick-up and drop-off location. The pick-up and drop-off locations observed between 6AM and 9AM on September 1, 2009 are displayed in Figure 3.

Overall, the average number of hourly trips throughout each day of the week is shown in Figure 4. The trips reveal two distinct patterns: weekdays and weekends. Weekday trips are seemingly driven by regular office and business hours as the trip count starts to rise around 8AM, drops down around 6PM, and jumps up slightly around 8PM (presumably post-dinner time when people travel back home or to another destination). It then gradually drops down and hits the lowest count around 4AM before rising again. The number of weekend trips does not fluctuate as much as the weekday trips. The busiest hour for the weekend trips is around 1PM, while the least active hour is late morning, around 8AM. The amount of weekend trips is higher than weekday trips from 1AM to 7AM but lower from 7AM to 9PM.

Once the pick-up and drop-off locations are determined, it is possible to connect all the pairs to draw up individual origin-destination trips. Collectively, these trips can be aggregated to generate citywide, area-based origin-destination movements. Unlike existing dynamic O-D matrices that are time-variant on fixed zonal areas [80]–[82], we tried to capture more realistic travel demand spatially as it emerged. This required an O-D matrix that is dynamic in both the temporal and spatial dimensions. To achieve this, we grouped individual origins (pick-ups) and destinations (drop-offs) according to their locations so that origins (or destinations) that were geographically close together could be clustered into

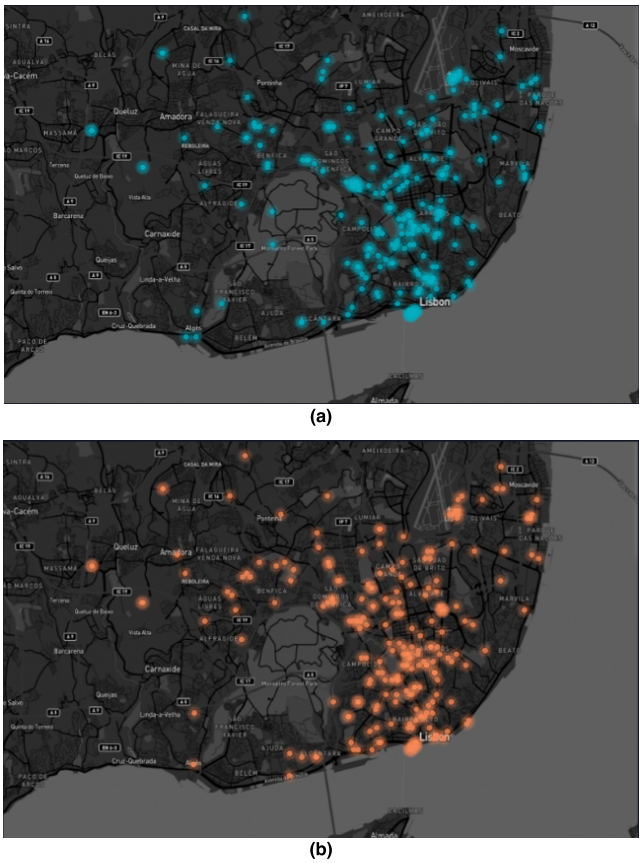


FIGURE 3. Pick-up locations (a) and drop-off locations (b) as determined from the taxi data based on service status information.

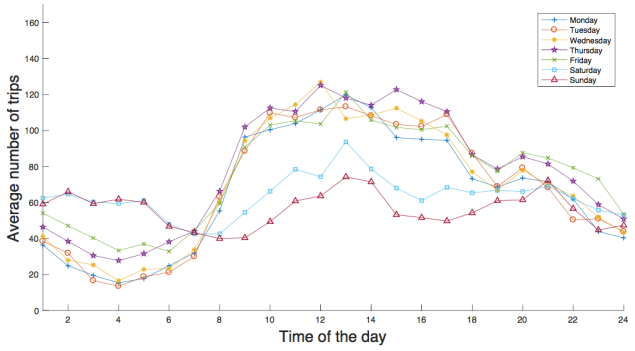


FIGURE 4. Average number of hourly trips on different days of the week.

an origin (or destination) area. Since the number of clustered origins or destinations is not known beforehand (there are no fixed taxi service stations such as taxi stands or taxi stops, for example), the clusters must gradually emerge based on the actual aggregate of trips. A clustering algorithm like k-means [84] cannot be used in this scenario because the number of clusters, or k, must be predefined for the clustering to proceed.

To rectify this shortcoming of k-means, we applied an approach called X-means clustering [85]. This approach can cluster data points without a predefined number of clusters since it estimates k by making local decisions about which subset of the current centroids should be split to better fit the data. Bayesian Information Criteria (BIC) [86] is used for the splitting decision, so the focus is on optimizing BIC value.

Once our taxi trajectory data was processed to identify pick-up and drop-off locations over a sliced period of interest, X-means clustering was performed for a separate set of pick-up and drop-off locations where each data point is a pair of geolocation coordinates to create clustering on a two-dimensional space. The X-means algorithm initializes k to 2 (to initially create two clusters). Each data point is iteratively assigned to the nearest centroid and the cluster centers are updated based on the renewed cluster mean. The data points are then reassigned and the cluster centers are updated again. This process continues to repeat. For each value of k, each centroid is split into two children, which are moved in opposite directions along a randomly-chosen vector for a distance proportional to the size of the region. A k-means algorithm is then run locally in each parent region with $k = 2$ for each pair of children. Data points are clustered to the children within the parent region. The split decision is then made locally based on the BIC value. The process continues iteratively until the global BIC value is optimized.

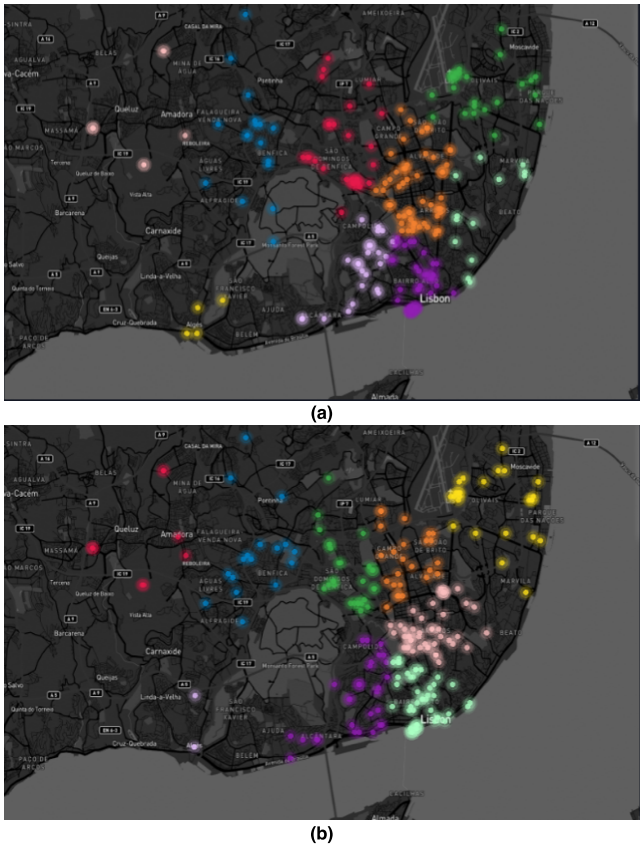


FIGURE 5. Clusters of pick-up and drop-off locations based on the X-means algorithm, with each cluster represented by a different color. (a) Clusters of pick-up locations. (b) Clusters of drop-off locations.

Figure 5 shows the nine clusters of pick-up locations and the nine clusters of drop-off locations that resulted from

applying the X-means algorithm to the observations during our period of interest. Clustering was done separately for the pick-up location data points and the drop-off location data points. It is a coincidence that this example yielded nine clusters for both pick-up and drop-off locations.

The X-means algorithm yielded origin and destination clusters. These clusters may contain noise or outliers such as data points that are widely spread out or that are notably distant from the cluster centroid. We filtered out these potential outliers by using the density-based spatial clustering of applications with noise (DBSCAN) algorithm [87], which groups data points together with many nearby neighbors and filters out outliers that are alone in low-density regions.

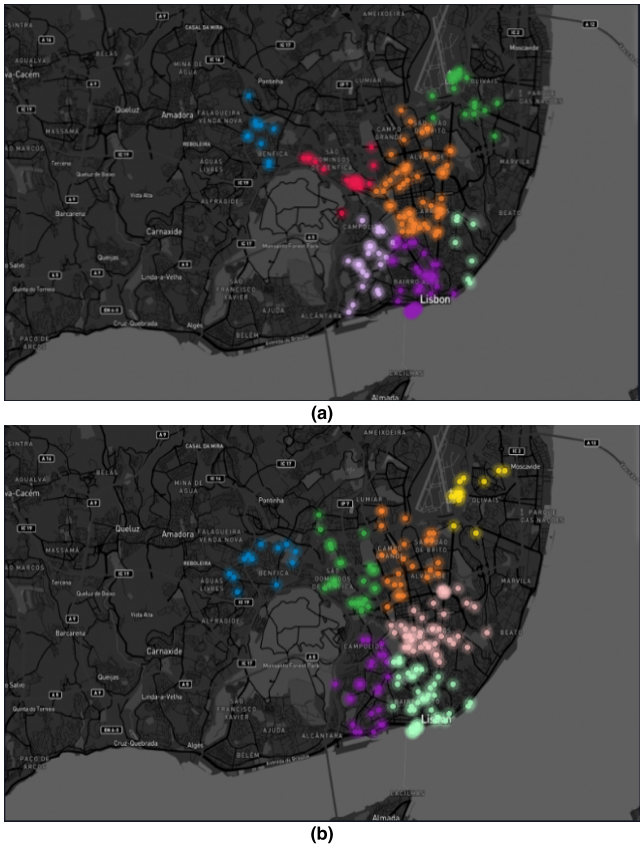


FIGURE 6. Outlier-filtered clusters of pick-up and drop-off locations based on the DBSCAN, with each cluster represented by a different color. (a) Filtered clusters of pick-up locations. (b) Filtered clusters of drop-off locations.

The DBSCAN algorithm requires two parameters: the maximum radius of the neighborhood (ε) and the minimum number of points required to form a dense region (*minPts*). We used $\varepsilon = 1,000$ m and $minPts = 5$, respectively. These parameter values were chosen and justified based on our observations of the results. Figure 6 shows the clusters of origins (a) and destinations (b) that resulted from applying the DBSCAN algorithm to filter out outliers. Different options for the DBSCAN parameters may be worth exploring in future studies.

Choosing optimal values for the two parameters is still an open research question, as shown by Wong and Huang [88] in their sensitivity analysis of spatiotemporal trajectory data clustering. The two parameters appear to be working against each other such that increasing the value of *minPts* not only reduces the number of clusters, but also the area that falls within each cluster. Increasing the value of ε produces extensive clusters. Although some studies have suggested appropriate values for these two parameters [89], [90], these recommendations were data-dependent and are not generally applicable.

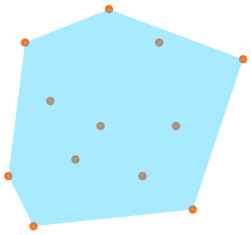


FIGURE 7. An origin area is represented with a convex polygon of the clustered pick-up locations.

To obtain area-based origin-destination flows for the construction of our O-D matrix, we needed to identify a geographical area for each cluster to represent each origin and destination area, which corresponds with each row and column of an O-D matrix. Each cluster is represented as an enclosed convex hull polygon [91] whose vertices are points from the cluster and which encompasses all of the cluster's points. Figure 7 shows an example polygon that represents an origin area where dots are the origins (or pick-up locations) and the convex polygon is the shaded area.

This approach enables the construction of an O-D matrix with $\boldsymbol{OD}_t = \left[T_{i,j}\right]_{N\times M}$ for observation period t, where $T_{i,j}$ is the trip volume from the origin i to destination j for $i = 1, 2, 3, \ldots, N$ and $j = 1, 2, 3, \ldots, M$. The matrix does not need to be square, with an equal number of origin and destination areas, as the origin and destination areas emerge dynamically with the actual travel demand. Figure 8 shows a time-dependent O-D matrix in the period between 6AM and 9AM on September 1, 2009 that was obtained using our approach. Travel demand, represented as the trip flow percentage from each of the seven origin areas to each destination area, is illustrated. Figure 9 shows a three-dimensional representation of trip flows from one origin area to all destination areas on the same map.

B. O-D MATRIX SIMILARITY MEASURE

Similar to other transportation systems, the demand for taxi services exhibits daily periodicity in time and space that reflects the patterns of underlying human activity [25]. An O-D matrix can capture demand variation on taxi services. A good taxi service provisioner should be able to cope with ever-changing demand. It is therefore important to understand the change, or difference, in the travel demands described by

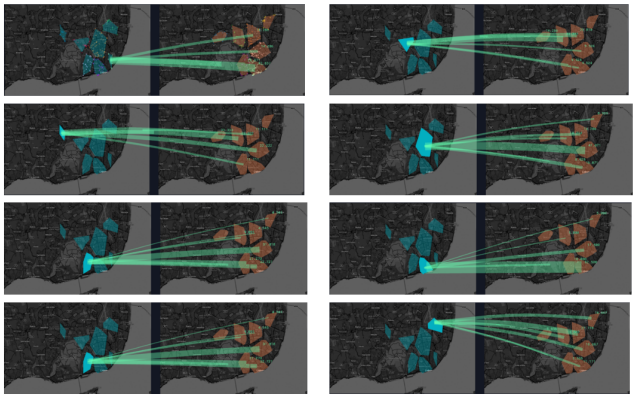


FIGURE 8. An example of trip flows based on a time-dependent O-D matrix constructed using the proposed approach for the period of 6AM - 9AM.

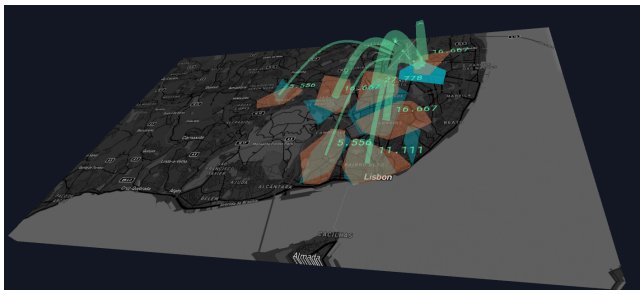


FIGURE 9. A 3D sample of trip flows from one origin area to all destination areas based on a time-dependent O-D matrix constructed for the period of 6AM - 9AM using the proposed approach, where the origin and destination areas are shown on the same map.

the O-D matrices over different periods to enable the effective management of taxi services operations.

Since our taxi O-D matrix varies spatially, its dimensions can change based on differing numbers of origin/destination area clusters to represent traffic analysis zones. Thus, existing approaches cannot be used to make comparisons between matrices. This eliminates methods such as matrix subtraction, R-squared, etc. We addressed this by developing a new approach to measure similarity between dynamic, spatiotemporal-variant O-D matrices.

The challenge in measuring similarity between matrices of different dimensions is transforming them into a comparable format such as *vectors* of the same size, whose similarity can simply be measured with *cosine similarity* as follows.

$$S(X,Y)=\cos(\theta)=\frac{\boldsymbol{X}\cdot\mathbf{Y}}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|}=\frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}X_i^2}\sqrt{\sum_{i=1}^{n}Y_i^2}}, \quad (1)$$

where X_i and Y_i are components of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The cosine similarity measures the similarity between two non-zero vectors of an inner product space by measuring the cosine of the angle between them. It is a comparison based on orientation but not magnitude. Two vectors with the same orientation (0 degrees apart) have a cosine similarity of 1. Two vectors that lie perpendicular to each other (oriented at 90 degrees) have a similarity of 0. Two vectors that are diametrically opposed (180 degrees apart) have a similarity of -1. This vector approach lets us transform the travel demand described by an O-D matrix into vectors and measure the similarity between two O-D matrices using cosine similarity and the approach described below.

Suppose that we want to measure the similarity between two O-D matrices, A and B. Matrix A has a total of N origins and M destinations while matrix B has a total of U origins and V destinations. Let O^A denote a set of the origins of A, i.e., $O^A=\{O_1^A, O_2^A, \ldots, O_N^A\}$ where each origin i contains the corresponding latitude and longitude of its centroid, $O_i^A = \{o_i^A\,(lat)\,, o_i^A\,(lon)\}$, and D^A denotes a set of destinations of A, i.e., $D^A = \{D_1^A, D_2^A, \ldots, D_M^A\}$ where each destination i contains the corresponding latitude and longitude of its centroid, $D_i^A = \{d_i^A\,(lat)\,, d_i^A\,(lon)\}$. Matrix B uses similar mathematical notation.

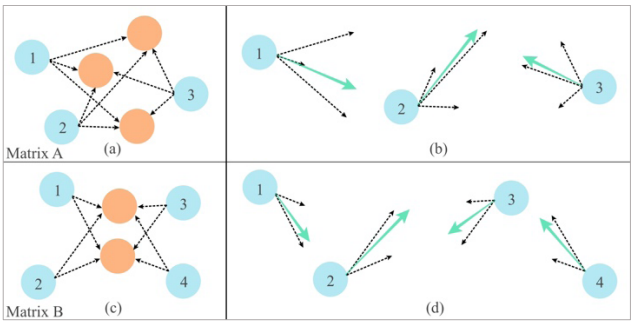


FIGURE 10. An example of resultant flows from three origins to three destinations in Matrix *A* and from four origins to two destinations in Matrix *B*: (a) Origins and flow directions to corresponding destinations in O-D Matrix *A*; (b) Flow directions (dashed lines) and resultant flow (solid line) of each origin in Matrix *A*; (c) Origins and flow directions to corresponding destinations in O-D Matrix *B*; (d) Flow directions (dashed lines) and resultant flow (solid line) of each origin in Matrix *B*.

Each origin i of A has travel demands that flow to each of the M destinations. These can be thought of as M vectors pointing from the origin i to each destination. A *resultant vector* [92] can be calculated as a sum of these M vectors to represent the overall flow with direction and magnitude, or the *resultant flow* from the origin i. Consequently, Matrix A will have N resultant flows originating from each of its N origins. Figure 10 shows the resultant flow of each origin of sample Matrices A and B, where A has three origins and three destinations and where B has four origins and two destinations.

For comparison, each resultant flow i can be written in the form of a vector where its vector components include characteristics of the result flow: direction, magnitude, origin location, and head location. Each resultant flow i of matrix A can be represented and written in a vector form as follows.

$$R_i^A = [\theta_i^A, T_i^A, O_i^A, H_i^A], \quad (2)$$

where θ_i^A is the direction (angle) of the resultant flow, T_i^A is the magnitude (the sum of travel demands), O_i^A is the

origin location, and H_i^A is the head location (the geolocation of the resultant vector's head, defined as $H_i^A = \{h_i^A\,(lat)\,, h_i^A\,(lon)\}$). Each of the derived U resultant flows of Matrix B can be written in the form of a vector, as shown in Eq. (2).

Although it is very unlikely, there is still a possibility that the sum of all outflows is zero, meaning that the flow's direction is zero due to the outflows' exact orientations. This could be the case, for example, with two outflows where one points to the east (0 degrees) and the other points to the west (180 degrees). The resulting flow's direction would be zero, indicating that the net displacement is zero because the outflow directions are precisely in opposite directions and canceled each other out. In a rare but possible scenario, a resultant flow X has a flow direction of zero ($\theta_i^X = 0$) due to east-west outflow orientations and another resultant flow Y also has a flow direction of zero ($\theta_i^Y = 0$) due to north-south outflow orientations, so both have a flow direction of zero. Differentiating between the two flows may not seem possible because of the loss of flow direction information. Although the flow direction is zero, however, the resultant flow still has three other dimensions (Eq. 2): its magnitude, its origin's geolocation, and its head's geolocation. This information can still convey other important characteristics about the flows and, to a certain extent, can be used to differentiate between two resultant flows. We nevertheless admit that the possibility of having a flow direction of zero due to exact orientations of the origin's outflows is one of the limitations of our approach, and is an area worth future study.

There are a total of N and U resultant flows for Matrices A and B, respectively. To measure the similarity between the matrices, the cosine similarity between R_i^A and R_j^B is calculated for $i = 1, 2, 3, \ldots, N$ and $j = 1, 2, 3, \ldots, U$ and the maximum value for each i and j, denoted by C_i^A and C_i^B respectively, is identified and kept for further calculation. The final *similarity* value (*Sim*) can then be calculated as the average of these pairwise-comparison maximum values as follows.

$$Sim = \frac{\sum_{i=1}^{N} C_i^A + \sum_{i=1}^{U} C_i^B}{N + U} \tag{3}$$

To summarize the approach in measuring the similarity between two time-dependent O-D matrices A and B of different dimensions, the algorithm below lists the steps.

Similarity between two O-D matrices (Eq. 3) is measured based on cosine similarities while accounting for the different characteristics of the flows: direction, magnitude, origin geolocation, and head geolocation. These characteristics are captured in a four-dimensional vector (Eq. 2). If the two O-D matrices are geographically nearby, the origin's geolocation elements are relatively alike but the other three elements (direction, magnitude, and head direction) may be different. Together, these all contribute to the similarity calculation. In a case where the two O-D matrices are geographically distant (e.g., different cities), all four elements contribute equally to the similarity calculation with a relatively larger gap in the origins' geolocation elements compared to the previous case. The physical meaning of the similarity measurement is thus the likeness of the flows' properties as characterized by four elements of the O-D matrices, which encapsulate the main characteristics in terms of direction, magnitude, and geography (origin and destination locations).

Algorithm 1 Time-Dependent O-D Matrix Similarity Measure

Input: Resultant flows of Matrices A and B ($\{R_i^A\}$ and $\{R_j^B\}$)
Output: Similarity value (*Sim*)

1. For $i \leftarrow 1$ to N (the number of origins of A) do
2. For $j \leftarrow 1$ to U (the number of origins of B) do
3. Compute cosine similarity $S(R_i^A, R_j^B)$
4. End
5. End
6. Determine the maximum value $C_i^A = \underset{j\in\{1,2,\ldots,U\}}{\text{argmax}} \left(S(R_i^A, R_j^B)\right)$ *For* $i \in \{1, 2, \ldots, N\}$
7. Determine the maximum value $C_j^B = \underset{i\in\{1,2,\ldots,N\}}{\text{argmax}} \left(S(R_i^A, R_j^B)\right)$ *For* $j \in \{1, 2, \ldots, U$
8. Compute $Sim = \frac{\sum_{i=1}^{N} C_i^A + \sum_{i=1}^{U} C_i^B}{N+U}$

V. RESULTS

Using the methodology described in the previous section, we measured the *cross-similarity* and *self-similarity* of taxi O-D flows to better understand the regularity and variation of taxi travel demand in Lisbon. Taxi trajectory data was used to construct an O-D matrix for each three-hour period of each day of the week. Three-hour periods begin at midnight (e.g., 0AM - 3AM, 3AM - 6AM, etc.), to create eight daily periods or a total 56 O-D matrices to represent the entire week's dynamic taxi travel demands.

Cross-similarity values between these 56 O-D matrices were measured. The result is shown in Figure 11, whose time-slots go from Monday 0AM – 3AM to Sunday 9PM – 0AM. Similarity was measured between each pair, starting from a comparison between the first time slot and the second time slot, then comparing the first and third time slots, and so on until the 56$^{\text{th}}$ time-slot. In total, we created $56 \times 56 = 3,136$ similarity values. Since the similarity measure is symmetrical (similarity between A and B is equal to similarity between B and A), the resulting matrix shown in Figure 11 is symmetric. Self-similarity yields the maximum value of 1, which is shown across the diagonal of Figure 11.

The resulting similarity values are relatively high (ranging from 0.8 to 1.0). This does not mean that the measurement cannot distinguish differences between the O-D matrices, but rather reflects on a reality that suggests most O-D matrices or travel demand patterns are quite similar. The implication is that travel demand is generally regular and shows recurrent patterns. This regularity in travel demand is favorably in line

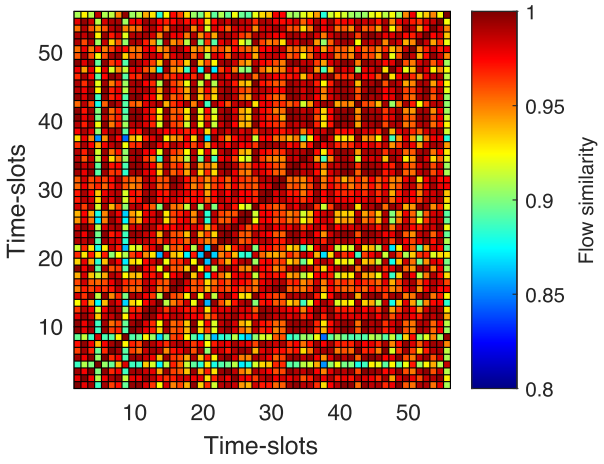


FIGURE 11. Cross-similarity result shown in a pseudocolor plot where each time slot goes from Monday 0AM – 3AM to Sunday 9AM – 0AM.

with previous studies of human mobility based on a massive mobile phone network data [23] and public bus transportation data [49].

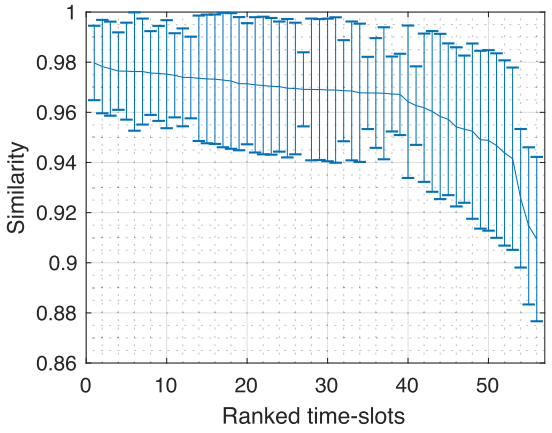


FIGURE 12. Cross-similarity result ranked from highest to lowest values according to the average over each time slot.

An average similarity value of each time slot in Figure 11 was calculated and ranked from the highest to lowest values. The results are shown in Figure 12, along with a respective standard deviation bar. The average similarity value gradually decreases and drops at a higher rate from the 40th-ranked time slot. It drops even more sharply at the 54th-ranked time slot.

The cross-similarity values measure how similar the taxi travel demand pattern in each time-slot is compared to other time slots. Effective taxi service operation management principles that are suitable for some time slots can be applied to others that exhibit similar patterns, and vice versa. Table 1 of Figure 12 shows the five time slots with the highest average similarity values, meaning they have the most common pattern with other time slots. Top time slots include Tuesday 6AM – 9AM, Sunday 9AM – 12AM, Wednesday 9PM – 0AM, Thursday 9AM – 12AM, and Sunday 3PM – 6PM. Table 2 of Figure 12 lists the bottom five time

TABLE 1. Top five time slots with highest average cross-similarity values.

Ranking	Time-slot	Cross-similarity value	Std. Dev.
1	Tue 6AM – 9AM	0.979702	0.014848
2	Sun 9AM – 12AM	0.978257	0.01863
3	Wed 9PM – 0AM	0.977426	0.018762
4	Thu 9AM – 12AM	0.976462	0.015445
5	Sun 3PM – 6PM	0.976433	0.019338

TABLE 2. Bottom five time slots with lowest average cross-similarity values.

Ranking	Time-slot	Cross-similarity value	Std. Dev.
56	Mon 9PM – 0AM	0.909448	0.032784
55	Mon 9AM – 12AM	0.914715	0.031344
54	Sun 6PM – 9PM	0.925724	0.027632
53	Tue 9AM – 12AM	0.941514	0.03636
52	Tue 12AM – 3PM	0.943798	0.036951

slots with the least common pattern: Monday 9PM – 0PM, Monday 9AM – 12AM, Sunday 6PM – 9PM, Tuesday 9AM – 12AM, and Tuesday 12AM – 3PM.

The results suggest that taxi travel demand that occurs within the top time slots can be interpreted as the most common pattern, while the demand that transpires during the bottom time slots can be thought of as the least common pattern. With the proposed similarity measurement, we were able to identify the most and least common time slots for taxi travel demand. This information is useful for taxi service operation management.

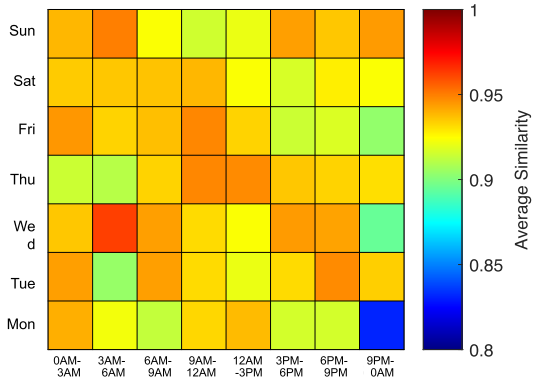


FIGURE 13. Self-similarity result shown in a pseudocolor plot.

We measured self-similarity by examining the similarity between taxi O-D matrices, or taxi travel demand occurring in the same time-slot over the course of two months. This allowed us to quantify the regularity of travel demand across different periods within a day. A higher self-similarity can imply a more regular pattern of taxi travel demand or more predictable taxi demand, while a lower self-similarity implies less predictability in demand. The self-similarity result is shown as an average value over two months for each time slot in Figure 13. Figure 14 shows the ranked values, along with the corresponding standard deviations.

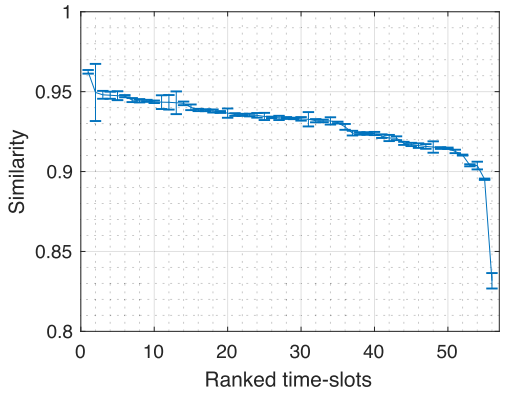


FIGURE 14. Self-similarity result ranked from the highest to lowest values.

TABLE 3. Top five time slots with highest average self-similarity values.

Ranking	Time-slot	Self-similarity value	Std. Dev.
1	Wed 6AM – 9AM	0.962412	0.001142
2	Sun 6AM – 9AM	0.949496	0.017876
3	Fri 9AM – 12AM	0.948017	0.002465
4	Thu 9AM – 12AM	0.947762	0.002221
5	Tue 6PM – 9PM	0.947445	0.002778

TABLE 4. Bottom five time slots with highest average self-similarity values.

Ranking	Time-slot	Self-similarity value	Std. Dev.
56	Mon 9PM – 0AM	0.831643	0.004832
55	Wed 9PM – 0AM	0.895183	0.000422
54	Fri 9PM – 0AM	0.903748	0.002414
53	Tue 3AM – 6AM	0.903916	0.000648
52	Thu 3AM – 6AM	0.910313	0.000302

The ranked self-similarity values drop sharply from the first to the second time slot. They then slowly decrease before dropping steeply between the 55^{th} and the last time slot. The top time slots are listed in Table 3 of Figure 12: Wednesday 6AM – 9AM, Sunday 6AM – 9AM, Friday 9AM – 12AM, Thursday 9AM – 12AM, and Tuesday 6PM – 9PM. The bottom time slots are listed in Table 4 of Figure 12: Monday 9PM – 0AM, Wednesday 9PM – 0AM, Friday 9PM – 0AM, Tuesday 3AM – 6AM, and Thursday 3AM – 6AM. These results suggest that travel demand in the top time slots is the most regular, while travel demand in the bottom time slots is the least regular compared to other periods.

This study has shown that taxi trajectory data can be used to develop spatiotemporal-varying O-D matrices. The self-similarities and cross-similarities of these matrices can be calculated. Our goal is to quantify the regularity of taxi travel demand across different time periods. The results show more taxi demand variation over a typical day than within the same time period on different days. This is true for both weekdays and weekends.

Our analysis opens new possibilities for gaining insight into how taxi trajectory data can be used to identify time periods that exhibit similarity and thus require similar operational strategies. This, in turn, will allow the development of improved demand-based taxi services. One problem with traditional taxi services is difficulty in matching taxi demand to supply when there is no phone booking or other reservation system. After a passenger drop-off in a given location and time, a taxi driver could be faced with a choice set comprising several zones for passenger pick-up. A location choice model for passenger pick-up can be developed to address this issue. The trip generation and attraction roles of each zone change based on the time of day, however, making it necessary to create location choice models for different times. Time-varying zone plans are structured to represent daily fluctuations of activity intensity in different parts of the city. Therefore, the time-varying zone plans developed for this study could inform the design of a reduced number of zones for modeling time-of-day location choices for passenger pick-up for taxi drivers.

Taxi trajectory data can also address the problematically low spatial and temporal resolutions of traditional surveys. GPS devices create the most accurate recordings of the times and positions of taxi movements. Using this data can improve trip-misrepresentation issues associated with self-reports of travels made by all modes. Trips inferred by taxi trajectory data are also geocoded and are not attached to any zoning system, allowing the data to be used at any level of aggregation so we can move away from standard zoning systems.

VI. DEMO

For demonstration purposes, a video clip showing how our analysis was carried out is available on YouTube at https://www.youtube.com/watch?v=IbcQWPf-W6I.

VII. CONCLUSION

An O-D matrix is an important input for transport models to assess new transport policies. It provides estimates of traffic volume between pairs of origin and destination zones based on relevant mobility data. With recent advances in information technology, opportunistic sensing datasets such as AVL, AFC, CDR, GPS data, and so on, have been used for O-D estimation in lieu of traditional travel surveys, which are expensive and time-consuming. AFC and AVL data from public transit system such as buses and trains does not provide actual passenger origin and destination information since these transportation modes operate on fixed schedules and stops. CDR data is sparse in time as it is only recorded when the device connects to the cellular network. It is also coarse in space because it is bounded by the level of cell tower density. Due to privacy concerns and regulations, large-scale GPS tracking data is limited.

The taxi industry has entered the era of technology by implementing onboard devices to assist taxi operations. Taxi trajectory data is one of the main datasets generated by these devices. Unlike the aforementioned data sources, taxi trajectory data can provide detailed origin and destination information. This study explored the use of taxi trajectory data collected from Lisbon, Portugal to construct a taxi

O-D matrix that is dynamic in both space and time. It also introduced a new measure of similarity between these dynamic O-D matrices.

The proposed approach for constructing a dynamic, time-dependent O-D matrix involves detecting pick-up and drop-off locations based on service status information, clustering the origins and destinations using the X-means algorithm, filtering out potential outliers using DBSCAN, and identifying groups of origins and destinations using convex hull. Unlike existing dynamic O-D matrices, which are time-variant on fixed zonal areas, our O-D matrix is spatially dynamic so that origin and destination zone sizes and locations are not fixed. Comparisons between matrices therefore cannot be made directly using traditional approaches like matrix subtraction, R-squared, GEH statistic, RMSE, and EBM since these measures all require the matrices to have identical dimensions. A new measure of similarity that uses the concept of a resultant vector to derive an O-D matrix consisting of resultant flows characterized by volumes (magnitudes) and directions (degree angles) is proposed. Cosine similarity is then used to measure the similarity from the resultant flows that have been converted to a vector form.

Our approaches allowed us to measure the cross-similarity and self-similarity of the taxi travel demand in Lisbon. The results revealed the periods in which the greatest and least common taxi travel demand occurred, as well as the periods in which the most- and least-regular travel demand patterns emerged This information is essential for informed taxi service operation management.

Our proposed approaches can be useful for constructing spatiotemporal-varying taxi O-D matrix and for measuring their similarity. We believe these similarities extend beyond the state-of-the-art O-D matrix estimation and similarity measures. The results regarding time-varying O-D flows and time-varying zone plans will inform the development of time-of-day location choice models and other dynamic models.

There were several limitations to our study. The first was the lack of ground truth validation of our approach. We didn't have actual travel demand information available to validate the construction method of our O-D matrix, but we believe that our approach is intuitive enough that it is valid and comparable to state-of-the-art methodologies. One issue with GPS trajectory data is lack of descriptors for key events that may occur during a trip. Fortunately, our data contains descriptors for key events such as service status (occupied, vacant). A transition of taxi meter status can be used to determine passenger pick-up and drop-off events and origin and destination locations. If there is lack of descriptors for key events during a trip, those events must then be inferred using data mining techniques. Previous research has shown that this can be achieved with reasonable accuracy [93]–[95].

The scope of this study was limited to constructing time-dependent O-D matrices with adaptive zoning schemes. Further research must be carried out to validate the results. Another potential limitation is the use of only 15% of the taxi fleet to characterize the taxi travel demand of the whole city. Future studies should attempt sample expansion to represent the mobility behaviour of the total taxi population in the study region. A third limitation is selecting optimal choices for the DBSCAN parameters, which is still an open research question that is worth future investigation. A final limitation relates to the possibility of the taxi data not reflecting on the actual location of the final destination because the drop-off location is different than the traveler's ultimate destination. In cases where the final destination is not the drop-off location, we believe that the final destination may not be too distant from the drop-off location. Passengers may continue their journey on foot or by bike for the final mile of travel. This so-called last mile travel is important and cannot be overlooked. We also did not calculate positional latency that causes deviation between the position outputted by the GPS and the vehicle's realtime position, which could potentially affect the precision of a trip's origin/destination. Further research must be conducted to investigate the validation method, to incorporate multi-source data, to consider the last mile travel, and to evaluate the impact of positional latency.

REFERENCES

[1] B. R. Hellinga, "Estimating dynamic origin-destination demands from link and probe counts," *Transp. Res. A*, vol. 31, no. 1, p. 83, 2002.
[2] H. Yang and H. Rakha, "A novel approach for estimation of dynamic from static origin–destination matrices," *Transp. Lett.*, vol. 11, no. 4, pp. 219–228, 2017.
[3] K. Ashok, "Estimation and prediction of time-dependent OD flows," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 1996.
[4] X. Zhou, S. Erdoğan, and H. Mahmassani, "Dynamic origin-destination trip demand estimation for subarea analysis," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1964, no. 1, pp. 176–184, 2006.
[5] E. Cascetta, *Teoria e Metodi dell'Ingegneria dei Sistemi di Trasporto*. UTET, 1998.
[6] H.-W. Chang, Y.-C. Tai, and J. Y.-J. Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Mining*, vol. 5, no. 1, p. 3, 2010.
[7] M. G. Demissie, S. Phithakkitnukoon, and L. Kattan, "Trip distribution modeling using mobile phone data: Emphasis on intra-zonal trips," *IEEE Trans. Intell. Transp. Syst.*, to be published.
[8] M. G. Demissie, S. Phithakkitnukoon, L. Kattan, and A. Farhan, "Understanding human mobility patterns in a developing country using mobile phone data," *Data Sci. J.*, vol. 18, no. 1, pp. 1–13, 2019.
[9] A. Hagen-Zanker and Y. Jin, "Improving geographic scalability of traffic assignment through adaptive zoning," in *Proc. Conf. Comput. Urban Planning Urban Manage.*, 2011, p. 15.
[10] A. Hagen-Zanker and Y. Jin, "A new method of adaptive zoning for spatial interaction models," *Geographical Anal.*, vol. 44, no. 4, pp. 281–301, 2012.
[11] H. Hammadou, I. Thomas, A. Verhetsel, and F. Witlox, "How to incorporate the spatial dimension in destination choice models: The case of Antwerp," *Transp. Planning Technol.*, vol. 31, no. 2, pp. 153–181, 2008.
[12] A. Hagen-Zanker and Y. Jin, "Adaptive zoning for transport mode choice modeling," *Trans. GIS*, vol. 17, no. 5, pp. 706–723, 2013.
[13] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*. Cambridge, MA, USA: MIT Press, 1987.
[14] M. Van Aerde, B. Hellinga, L. Yu, and H. Rakha, "Vehicle probes as real-time ATMS sources of dynamic OD and travel time data," in *Proc. Adv. Traffic Manag. Conf. Large Urban Syst.*, 1993, pp. 207–230.
[15] L. Kattan and B. Abdulhai, "Sensitivity analysis of an evolutionary-based time-dependent origin/destination estimation framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1442–1453, Sep. 2012.
[16] X. Yang, Y. Lu, and W. Hao, "Origin-destination estimation using probe vehicle trajectory and link counts," *J. Adv. Transp.*, vol. 2017, Jan. 2017, Art. no. 4341532.

[17] X. Zhou and H. S. Mahmassani, "Dynamic origin-destination demand estimation using automatic vehicle identification data," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 105–114, Mar. 2006.

[18] J. Sun and Y. Feng, "A novel OD estimation method based on automatic vehicle identification data," in *Proc. Int. Conf. Intell. Comput. Inf. Sci.*, 2011, pp. 461–470.

[19] C. Antoniou, M. Ben-Akiva, and H. N. Koutsopoulos, "Dynamic traffic demand prediction using conventional and emerging data sources," *IEE Proc.-Intell. Transp. Syst.*, vol. 153, no. 1, pp. 97–104, Mar. 2006.

[20] M. G. Mcnally, "The four step model," in *Handbook of Transport Modelling*. Irvine, CA, USA: Univ. California Irvine, 2007.

[21] V. Marzano, A. Papola, and F. Simonelli, "Limits and perspectives of effective O–D matrix correction using traffic counts," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 2, pp. 120–132, 2009.

[22] H. Yang, Y. Iida, and T. Sasaki, "An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts," *Transp. Res. B, Methodol.*, vol. 25, no. 5, pp. 351–363, 1991.

[23] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[24] C. Yang, F. Yan, and S. V. Ukkusuri, "Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system," *Transportmetrica A, Transp. Sci.*, vol. 14, no. 7, pp. 576–597, 2018.

[25] L. Moreira-Matias, J. Gama, M. Ferreira, and L. Damas, "A predictive model for the passenger demand on a taxi network," in *Proc. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2012, pp. 1014–1019.

[26] H.-W. Chang, Y.-C. Tai, H.-W. Chen, J. Y.-J. Hsu, and C. P. Kuo, "iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches," in *Proc. TAAI*, 2008, pp. 1–8.

[27] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 109–118.

[28] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Advances in Spatial and Temporal Databases* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Germany: Springer, 2011.

[29] X. Wan, J. Kang, M. Gao, and J. Zhao, "Taxi origin-destination areas of interest discovering based on functional region division," in *Proc. 3rd Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2013, pp. 365–370.

[30] J. Lee, I. Shin, and G.-L. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Proc. 4th Int. Conf. Netw. Comput. Adv. Inf. Manage. (NCM)*, Sep. 2008, pp. 199–204.

[31] A. Lacombe and C. Morency, "Modeling taxi trip generation using GPS data: The Montreal case," in *Proc. Transp. Res. Board 95th Annu. Meeting*, 2016, Paper 16-4345.

[32] C. Yang and E. J. Gonzales, "Modeling taxi trip demand by time of day in New York city," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2429, no. 1, pp. 110–120, 2014.

[33] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces," in *Proc. UbiComp*, 2012, pp. 943–949.

[34] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *J. Geograph. Syst.*, vol. 14, no. 4, pp. 463–483, 2012.

[35] J. Zhu and X. Ye, "Development of destination choice model with pairwise district-level constants using taxi GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 93, pp. 410–424, Aug. 2018.

[36] J. Tang, S. Zhang, X. Chen, F. Liu, and Y. Zou, "Taxi trips distribution modeling based on entropy-maximizing theory: A case study in Harbin City—China," *Phys. A, Stat. Mech. Appl.*, vol. 493, pp. 430–443, Mar. 2018.

[37] M. Veloso, S. Phithakkitnukoon, C. Bento, N. Fonseca, and P. Olivier, "Exploratory study of urban flow using taxi traces," in *Proc. 1st Workshop Pervasive Urban Appl. (PURBA)*, San Francisco, CA, USA, 2011, pp. 1–8.

[38] M. Batty, "Thinking about cities as spatial events," *Environ. Planning B, Planning Des.*, vol. 29, pp. 1–2, Feb. 2002.

[39] A. Bertaud, "The Spatial organization of cities: Deliberate outcome or unforeseen consequence?" Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep., 2004.

[40] M. G. Demissie, "Combining datasets from multiple sources for urban and transportation planning: Emphasis on cellular network data," Civil Eng. Dept., Univ. Coimbra, Coimbra, Portugal, Tech. Rep., 2014.

[41] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer, "Mobile landscapes: Graz in real time," *Location Based Services and TeleCartography*. Berlin, Germany: Springer, 2005, pp. 433–444.

[42] Y. Jin and I. N. Williams, "A new land use and transport interaction model for London and its surrounding regions," in *Proc. Eur. Transp. Conf.*, 2002.

[43] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape Urban Planning*, vol. 106, no. 1, pp. 73–87, 2012.

[44] M. V. Chitturi, J. W. Shaw, J. R. Campbell, and D. A. Noyce, "Validation of origin–destination data from bluetooth reidentification and aerial observation," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2430, no. 1, pp. 116–123, 2014.

[45] X. Huang, M. Ghodsi, and H. Hassani, "A Novel similarity measure based on eigenvalue distribution," *Trans. A. Razmadze Math. Inst.*, vol. 170, no. 3, pp. 352–362, 2016.

[46] R. Horn and C. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[47] S. Bhattacharya, S. Phithakkitnukoo, P. Nurmi, A. Klami, M. Veloso, and C. Bento, "Gaussian process-based predictive modeling for bus ridership," in *Proc. ACM Conf. Pervasive Ubiquitous Comput. Adjunct Publication UbiComp*, 2013, pp. 1189–1198.

[48] S. Foell, S. Phithakkitnukoon, G. Kortuem, M. Veloso, and C. Bento, "Predictability of public transport usage: A study of bus rides in Lisbon, Portugal," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2955–2960, Oct. 2015.

[49] S. Foell, S. Phithakkitnukoon, M. Veloso, G. Kortuem, and C. Bento, "Regularity of public transport usage: A case study of bus rides in Lisbon, Portugal," *J. Public Transp.*, vol. 19, no. 4, p. 10, 2016.

[50] C. Somdulyawat, P. Pongjitpak, S. Phithakkitnukoon, M. Veloso, and C. Bento, "A tool for exploratory visualization of bus mobility and ridership: A case study of Lisbon, Portugal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. UbiComp*, 2015, pp. 1117–1121.

[51] N. Caceres, J. P. Wideberg, and F. G. Benitez, "Review of traffic data estimations extracted from cellular networks," *IET Intell. Transp. Syst.*, vol. 2, no. 3, pp. 179–192, Sep. 2008.

[52] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

[53] M. G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, and C. Bento, "Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: A case study of Senegal," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2466–2478, Sep. 2016.

[54] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel," *Transp. Res. C, Emerg. Technol.*, vol. 15, no. 6, pp. 380–391, 2007.

[55] M. G. Demissie, G. H. de Almeida Correia, and C. Bento, "Intelligent road traffic status detection system through cellular networks handover information: An exploratory study," *Transp. Res. C, Emerg. Technol.*, vol. 32, pp. 76–88, Jul. 2013.

[56] H. X. Liu, A. Danczyk, R. Brewer, and R. Starr, "Evaluation of cell phone traffic data in minnesota," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2086, no. 1, pp. 1–7, 2008.

[57] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PLoS One*, vol. 7, no. 6, 2012, Art. no. e39253.

[58] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Apr. 2011.

[59] S. Phithakkitnukoon, T. Sukhvibul, M. Demissie, Z. Smoreda, J. Natwichai, and C. Bento, "Inferring social influence in transport mode choice using mobile phone data," *EPJ Data Sci.*, vol. 6, no. 1, p. 11, 2017.

[60] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring Land Use from Mobile Phone Activity," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–8.

[61] M. G. Demissie, G. Correia, and C. Bento, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage," *Transportmetrica A, Transp. Sci.*, vol. 11, no. 6, pp. 502–524, 2015.

[62] Y. Lu and Y. Liu, "Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 105–108, 2012.

[63] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Sci.*, vol. 7, no. 1, p. 2, 2018.

[64] S. Phithakkitnukoon, T. Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, and R. Shibasaki, "Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan," *Pervasive Mobile Comput.*, vol. 18, pp. 18–39, Apr. 2015.

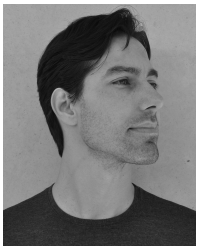
[65] R. Ásmundsdóttir, "Dynamic OD matrix estimation using floating car data," Centre Transp. Navigat., Delft Univ. Technol., Delft, The Netherlands, Tech. Rep., 2008.
[66] Y. Yang, H.-P. Lu, and Q.-Z. Hu, "A bi-level programming model for origin-destination estimation based on FCD," in *Proc. 10th Int. Conf. Chin. Transp. Prof. (ICCTP)*, 2010, pp. 117–124.
[67] S. M. Eisenman and G. F. List, "Using probe data to estimate OD matrices," in *Proc. 7th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2004, pp. 291–296.
[68] P. Cao, T. Miwa, T. Yamamoto, and T. Morikawa, "Bilevel generalized least squares estimation of dynamic origin–destination matrix for urban network with probe vehicle data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2333, no. 1, pp. 66–73, 2013.
[69] T. Guozhen, L. Lidong, W. Fan, and W. Yaodong, "Dynamic OD estimation using Automatic Vehicle Location information," in *Proc. 6th IEEE Joint Int. Inf. Technol. Artif. Intell. Conf.*, Aug. 2011, pp. 352–355.
[70] C. Basnayake, *Automated Traffic Incident Detection Using GPS Based Transit Probe Vehicles*. Calgary, AB, Canada: Calgary Univ., 2004.
[71] A. Sbaï, H. J. van Zuylen, J. Li, F. Zheng, and F. Ghadi, "Estimation of an urban OD matrix using different information sources," in *Computational Science and Its Applications* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Germany: Springer, 2017.
[72] P. S. Castro, D. Zhang, and S. Li, "Urban traffic modelling and prediction using large scale taxi GPS traces," in *Pervasive Computing* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence Lecture Notes Bioinformatics). New York, NY, USA: ACM, 2012.
[73] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining-KDD*, 2011, pp. 316–324.
[74] K. Liu, T. Yamamoto, and T. Morikawa, "Feasibility of using taxi dispatch system as probes for collecting traffic Information," *J. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 16–27, 2009.
[75] J. Tang, J. Liang, S. Zhang, H. Huang, and F. Liu, "Inferring driving trajectories based on probabilistic model from large scale taxi GPS data," *Phys. A, Stat. Mech. Appl.*, vol. 506, pp. 566–577, Sep. 2018.
[76] T. Horanont, S. Phithakkitnukoon, T. W. Leong, Y. Sekimoto, and R. Shibasaki, "Weather effects on the patterns of people's everyday activities: A study using GPS traces of mobile phone users," *PLoS ONE*, vol. 8, no. 12, 2013, Art. no. e81153.
[77] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Sensing urban mobility with taxi flow," in *Proc. 3rd ACM SIGSPATIAL Int. Workshop Location-Based Social Netw. (LBSN)*, 2011, pp. 41–44.
[78] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, 2013, Art. no. 17.
[79] P. Prommaharaj, S. Phithakkitnukoon, M. Veloso, and C. Bento, "Visualization tool for taxi usage analysis: A case study of Lisbon, Portugal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. UbiComp*, 2016, pp. 1343–1348.
[80] T. Toledo and T. Kolechkina, "Estimation of dynamic origin–destination matrices using linear assignment matrix approximations," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 618–626, Jun. 2013.
[81] L. Montero, E. Codina, and J. Barceló, "Dynamic OD transit matrix estimation: Formulation and model-building environment," in *Progress in Systems Engineering* (Advances in Intelligent Systems and Computing). Berlin, Germany: Springer, 2015.
[82] C.-C. Lu, X. Zhou, and K. Zhang, "Dynamic origin–destination demand flow estimation under congested traffic conditions," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 16–37, Sep. 2013.
[83] R. Darbéra, "Taxicab regulation and urban residents' use and perception of taxi services: A survey in eight cities," in *Proc. 12th WCTR*, Jul. 2010, Art. no. 01536.
[84] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: ACM, 1973.
[85] D. Pelleg, D. Pelleg, A. W. Moore, and A. W. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. Mach. Learn. table contents*, 2000, pp. 727–734.
[86] R. E. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *J. Amer. Stat. Assoc.*, vol. 90, no. 431, pp. 928–934, 1995.
[87] M. Daszykowski and B. Walczak, "Density-based clustering methods," in *Comprehensive Chemometrics*. Amsterdam, The Netherlands: Elsevier, 2010.
[88] D. W. S. Wong and Q. Huang, "Sensitivity of DBSCAN in identifying activity zones using online footprints," in *Proc. Spatial Accuracy*, 2016, pp. 151–156.
[89] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1996, pp. 226–231.
[90] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: An interactive clustering approach," in *Proc. 12th Annu. ACM Int. Workshop Geographic Inf. Syst.*, 2004, pp. 266–273.
[91] R. L. Graham, "An efficient algorith for determining the convex hull of a finite planar set," *Inf. Process. Lett.*, vol. 1, no. 4, pp. 132–133, 1972.
[92] P. M. Cohn, *Elements of Linear Algebra*. Oxfordshire, U.K.: Routledge, 2017.
[93] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in The Netherlands," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 3, pp. 285–297, 2009.
[94] H. Gong, C. Chen, E. Bialostozky, and C. T. Lawson, "A GPS/GIS method for travel mode detection in New York City," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 131–139, 2012.
[95] J. Du and L. Aultman-Hall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues," *Transp. Res. A, Policy Pract.*, vol. 41, no. 3, pp. 220–232, 2007.

WERABHAT MUNGTHANYA received the B.Sc. degree in statistics from Chiang Mai University, Thailand. He is currently a graduate student in the Department of Computer Engineering, Chiang Mai University. His research interests include data science, intelligent transportation systems, and visual analytics.

SANTI PHITHAKKITNUKOON received the B.S. and M.S. degrees in electrical engineering from Southern Methodist University, USA, in 2003 and 2005, respectively. He received the Ph.D. degree in computer science and engineering from the University of North Texas, USA. He is currently an Associate Professor in the Department of Computer Engineering, Chiang Mai University, Thailand. Before joining Chiang Mai University, he was a Lecturer in Computing at The Open University, U.K., a Research Associate at Newcastle University, U.K., and a Postdoctoral Fellow at the SENSEable City Laboratory of the Massachusetts Institute of Technology, USA. His research interest include urban informatics.

MERKEBE GETACHEW DEMISSIE received the M.Sc. degree in transport systems from the Royal Institute of Technology (KTH), Sweden, in 2009, and the Ph.D. degree in transportation systems from the MIT-Portugal program, in 2014. He is currently a Postdoctoral Associate with the Department of Civil Engineering, University of Calgary, Canada. Before joining the University of Calgary, he was a Postdoctoral Fellow with the University of Coimbra and with the Instituto Pedro Nunes, Portugal. His main research interests include transport demand modeling, intelligent transportation systems, data mining, and machine learning.

LINA KATTAN received the Ph.D. degree in civil engineering from the University of Toronto, Canada. She is a Professor of civil engineering with the University of Calgary, Canada. She also holds an Urban Alliance Professorship in Transportation Systems Optimization. Her research interests include advanced traffic management and information systems, including Intelligent Transportation Systems (ITS), traffic control, the application of artificial intelligence to ITS, connected and autonomous vehicles, network microsimulation modeling and analysis, dynamic traffic assignment, dynamic demand modeling, and traveler behavioral modeling in response to traffic and transit information.



MARCO VELOSO received the B.S., M.S., and Ph.D. degrees in Informatics Engineering from the University of Coimbra. He is an Adjunct Professor with the Polytechnic Institute of Coimbra, Portugal and a Researcher with the Center for Informatics and Systems of University of Coimbra, where he is a member of the Ambient Intelligence Laboratory. His research interest includes the use of data mining techniques on big data for smart cities and Intelligent Transportation Systems (ITS).

CARLOS BENTO received the Ph.D. degree in informatics engineering from the University of Coimbra, Portugal. He is currently an Associate Professor in habilitation with the University of Coimbra, where he is the Director of the Ambient Intelligence Laboratory, Centre for Informatics and Systems, and the Director of the Laboratory on Informatic Systems, Instituto Pedro Nunes (IPN), Coimbra. He has over 100 publications comprising papers in international journals and conferences and book chapters. His research has recently addressed the role of urban data on improving decisions for better quality of life in urban areas.

CARLO RATTI received the Ph.D. degree in architecture from the University of Cambridge. He is an architect and engineer who practices architecture in Turin and teaches at MIT, where he directs the SENSEable City Laboratory. His research interests include urban design, human -computer interfaces, electronic media, and the design of public spaces. He is a member of the Ordine degli Ingegneri di Torino, the Architects Registration Board (U.K.), and the Association des Anciens Eléves de l'École Nationale des Ponts et Chaussées.