REGULAR ARTICLE   **Open Access**

Understanding vehicular routing behavior with location-based service data

Yanyan Xu[1,2,3,4*] (iD), Riccardo Di Clemente[2,5,6] and Marta C. González[2,3,4,7]

*Correspondence:
yanyanxu@berkeley.edu
[1] MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, 200240, Shanghai, China
[2] Department of Civil & Environmental Engineering, MIT, 02139, Cambridge, MA, USA
Full list of author information is available at the end of the article

Abstract

Properly extracting patterns of individual mobility with high resolution data sources such as the one extracted from smartphone applications offers important opportunities. Potential opportunities not offered by call detailed records (CDRs), which offer resolutions triangulated from antennas, are route choices, travel modes detection and close encounters. Nowadays, there is not a standard and large scale data set collected over long periods that allows us to characterize these. In this work we thoroughly examine the use of data from smartphone applications, also referred to as location-based services (LBS) data, to extract and understand the vehicular route choice behavior. Taking the Dallas-Fort Worth metroplex as an example, we first extract the vehicular trips with simple rules and reconstruct the origin-destination matrix by coupling the extracted vehicular trips of the active LBS users and the United States census data. We then present a method to derive the commonly used routes by individuals from the LBS traces with varying sample rate intervals. We further inspect the relation between the number of routes and the trip characteristics, including the departure time, trip length and travel time. Specifically, we consider the travel time index and buffer index for the LBS users taking different number of routes. Empirical results demonstrate that during the peak hours, travelers tend to reduce the impact of traffic congestion by taking alternative routes. Overall, the proposed data analysis framework is cost-effective to treat sparse data generated from the use of smartphones to inform routing behavior. The potential in practice is to inform demand management strategies, by targeting individual users while generating large scale estimates of congestion mitigation.

Keywords: Routing behavior; Route detection; Location-based services; Origin-destination matrix; Traffic congestion

1 Introduction

With the growing population in cities and the restructuring of urban economies and societies, a fundamental task of transportation planners and engineers is to effectively move people and goods [1]. However, due to the daily increasing contradiction between the travel demand of citizens and the limit of road resources, the worsening traffic congestion not only causes tremendous economic loss and environmental problems but also has profound impacts on the public health [2–4]. Today, a number of strategies have been proposed by policymakers and researchers over the world to relieve the traffic congestion,

Springer

including advanced traffic signal control [5], strengthening the public transportation [6], reducing the travel demand of private vehicles [7], route recommendations [8], congestion pricing [9], or even looking into the future on autonomous vehicles [10].

Among the varying travel management strategies, understanding human mobility is always a fundamental task, supporting advanced decision systems. Thanks to the development of modern information and communication technologies (ICT) and the high penetration rate of mobile phone devices, researches can leverage on a large amount of digital traces with time stamps and geographical locations to understand and reproduce human mobility [11–13]. Despite daily destinations of human mobility can be modeled at half square kilometer scale, the routing behavior of travelers in their road networks is still not well modeled from ICT data sources [14]. A general solution to the lack of complete information is leveraging traffic assignment models, such as the user equilibrium assignment, dynamic traffic assignment, and the multi-agent approach, to assign each traveler to specific routes [15]. These models assume that travelers choose their routes with the intention of minimizing travel costs, such as travel distance or time. To this end, individuals are assumed to have perfect or partial information about the alternatives available to them. However, the choice of route is simultaneously affected by multiple factors, and the route choice behavior of people follows the bounded rationality principle [16]. This means that travelers can neither find the optimal routes because of the lack of accurate information about the traffic conditions, nor willing to spend much effort to obtain the optimized decision from complicated situations [17]. With vehicular trajectories collected periodically from hundreds of travelers during several months, starting when the driver turns the engine on, until it is turned off, Lima et al. found that the individual routing behavior is independent of the urban layout. People always have a dominant route and the alternative routes are bounded within an elliptic shape of high eccentricity [18]. In this work, we attempt similar analysis with the additional challenge imposed by data with less temporal accuracy and much lower frequency of collection, but much more pervasive than the vehicular trajectory data.

The prevailing big data resources utilized to model travel behavior include mobile phone data, check-in data at places of interest, and data collected by transportation agencies, like the floating car data. Among them, the mobile phone data, a.k.a. call detail records (CDRs), are passively collected and have the largest coverage. CDRs have been used to understand the mobility behavior [19], and reproduce the aggregated travel demand [20] and the trips chains (a sequence of visited places with timestamps) of the population [13, 21, 22]. However, methods solely relying on CDRs can not infer the routes taken by individuals due to coarse spatial resolution. As one kind of location-based services data, the geo-tagged check-in data provide more accurate locations than CDRs, but the data can be collected only when the users actively "checked-in" at their places of interest, making it a too infrequent source. Such data is activity-dependent and can not continuously record the traces of users [23]. Floating car data (FCD) are collected by transportation agencies for some specific purposes, recording the locations and speeds of floating cars. The GPS trajectories of taxis are the most commonly used FCD to analyze the traffic states. However, because of the cruising behavior of taxi drivers to search potential passengers, the taxis' trajectories are not perfect to study the route choice behavior of residents. In this work, we explore the location-based services (LBS) data, which specifically refer to the collection of the check-in or trajectory data generated by a set of smartphone applications. LBS use a

smartphone's localization technology (i.e. GPS, Wi-Fi) to track the holder's location down to a street address, if the holder has opted-in to allow the service to do that. Compared to the check-in data from one single application, our LBS data collect the locations from multiple applications and have much higher sampling frequency for two reasons, (i) some applications continuously collect the locations for providing map-related services to users; (ii) the aggregation of records from multiple applications also increase the sampling frequency. In recent years, LBS data have been used to examining meaningful visited places and social mixing [24, 25], travel behavior mining [26], and commuting pattern estimation [27]. The emergence of data collaborative using LBS records in the light of COVID19 pandemic has accelerated their use and value [28, 29].

This paper aims at analyzing the LBS for urban scale mobility demand, with focus in gaining insights on their use to extract routing behavior. Focusing on the Dallas-Fort Worth (DFW) metroplex, we first describe the process to impute vehicular trips from LBS data and then present a framework to deal with sparse data. Utilizing LBS data to analyze routing behavior faces important challenges: (i) LBS data are collected when the applications are activated, whenever the users are staying in one place or moving in unknown travel modes; (ii) LBS data are collected with varying sample rates, which hinder the detection of actual routes. We present detailed steps to resolve these issues. Further, we analyze the change of routing behavior by connecting it with the number of trips, the travel distance, and the travel time during peak hours. The main contributions of this work are summarized as follows: (i) we present detailed steps to process the LBS data, extract the vehicular trios and detect routes without the use of the road network and map-matching; (ii) we analyze differences of travelers' routing behavior by different number of trips, travel distances, and the time periods in one day. (iii) we inspect the impact of traffic congestion on individuals' routing behavior using two metrics, the travel time index and buffer index. Empirical results confirm that individuals that explore more routes can reduce the impact of congestion and increase their reliability of travel times. The complete implementation of all of our data analysis framework can be found at https://github.com/humnetlab/RoutingBehavior.

The rest of the paper is organized as follows. In Sect. 2, we give an overview of the LBS data used in this work; Sect. 3 depicts the methodology to process the LBS data and find the travelers' routes; Sect. 4 analyzes the route choice behavior and its connection to multiple factors. Finally, we conclude the work in Sect. 5.
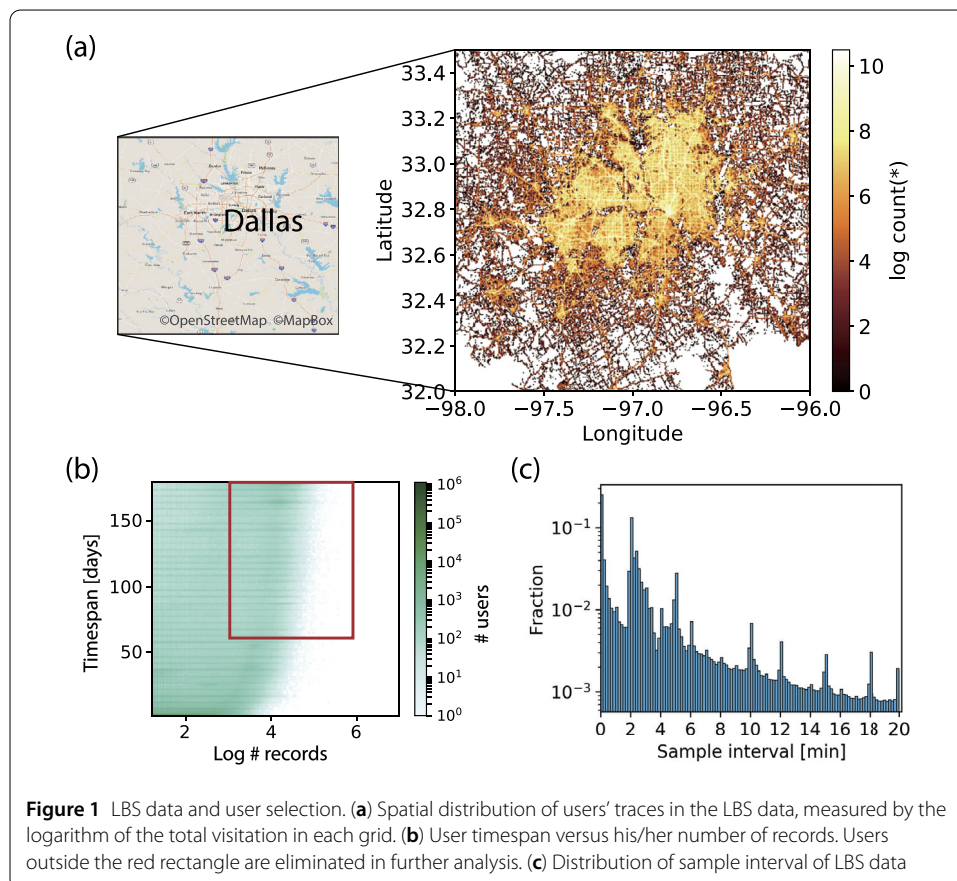
2 Data description

LBS are services offered to users through applications installed on smart mobile devices. Geographical locations of the users are simultaneously and actively collected by the application developers or map service operators. The users are normally positioned by global positioning system (GPS) or Wi-Fi positioning system (WPS), which are fairly accurate in space and offer new opportunities to study human activity and its complex interaction with the built environment at fine scale [30–34].

The LBS data used in this work are provided by Cuebiq, a location intelligence and measurement platform [35]. The datasets cover the DFW metroplex in Texas and were collected over a period of 6 months, from November 1st, 2016 to April 30th, 2017. The total number of users is approximate to 6.5 million and these users generated about 12.43 billion records in the given region and time period. Each LBS record consists of the

pseudonymized user ID, timestamp and geographical coordinate. Figure 1(a) illustrates the covered region and the visitation count in each grid cell in the first week of November 2016. The entire region is divided into 512×512 cells with approximate size 360×320 m^2, down to the block level. The highlighting of freeways and downtown in the heatmap indicates that they are the busiest places in terms of visitation counts.

As LBS data are collected when the user is interacting with the application, the collection can be interrupted if the user stops using the application. Besides, the applications are used with variant frequency. Thus, the users in LBS data have different numbers of records and timespan, which is defined as the time difference between the first and last records of the user. In Fig. 1(b), we show the timespan versus the number of records for all users in a heatmap. The region with dark green indicates that a large number of users are associated with the corresponding timespan and the number of records. As we can observe, a considerable part of users were recorded during a long term but have small numbers of records as they are not using the applications frequently. Other users have a considerable number of records but short timespan. These might be temporary visitors to the DFW metroplex or short time adopters of the app. For exploring the routing behavior, we require long-term observation of moving traces. To that end, we select the users whose data are collected over 60 days or more and have more than 1000 records, as enclosed by a red rectangle in Fig. 1(b). As a result, 13% of the users and 86% of the records are kept for further analysis.

Figure 1 LBS data and user selection. (**a**) Spatial distribution of users' traces in the LBS data, measured by the logarithm of the total visitation in each grid. (**b**) User timespan versus his/her number of records. Users outside the red rectangle are eliminated in further analysis. (**c**) Distribution of sample interval of LBS data

An important challenge remains, even with this sample, the records of LBS data are collected with variant, usually low frequency, because of the intermittent use of applications and different sample rates of applications. The LBS datasets are collected by a number of mobile applications (Apps) when the mobile phone user is interacting with these Apps or keeps these Apps running in the background. The sample interval, defined as the time difference between two consecutive records of the same user, is not fixed. Figure 1(c) shows the distribution of sample intervals in the LBS data. The sample intervals of a large proportion of records are larger than 2 min, which much lower frequency than floating car data and hinder the routing behavior analysis, especially in dense road networks. Besides, the uneven sample rates shown in Fig. 1(c) are mainly caused by the aggregation of records from multiple applications. Next, we propose a method to deal with this limitation and extract some valuable information from this data source.

3 Methodology

For analyzing the route choice behavior, a primary task is to map the users to specific routes they were taking. However, tracking the routes from LBS data is a challenge in the following two aspects: (i) LBS collect the data of users when they stay or move with all kinds of travel modes, e.g., walking, biking, driving and public transportation. Vehicular trips must first be imputed from the raw data for further route choice behavior analysis; (ii) LBS data are collected from multiple applications with different sample rates and at low-resolution. This hinders the entire routes over the road networks. Once the vehicular trips are assigned, we then select high-resolution trips for route detection and find the routes of other low-resolution trips by aligning them with the high-resolution ones.

3.1 Vehicular trips detection

For the records of each user, we first partition her records into a sequence of trips by looking into the time difference between two consecutive records. After the user selection illustrated in Fig. 1(b), the remaining users are labeled as high-frequency ones, a.k.a, active users. In this context, we suppose that a user starts a new trip if there is no record for at least 30 minutes before the current one, that is, $t_{current} - t_{previous} \geq 30$ min. Then we drop out the trips with less than 5 records. At this point, the records of each user have been partitioned into a sequence of trips in all kinds of travel modes.

A number of methods have been proposed to derive the travel mode from trajectory data, most of them process the high-resolution GPS traces or utilize sophisticated learning methods that require gold labels of travel modes for model training [36, 37]. Here we use a simple rule to identify the trip as a vehicular trip if its average speed is between 20 km/h and 100 km/h, leaving room for further improvements. In addition, there are trips with outliers caused by the GPS drift, which are eliminated in our experiments. We label the points which have less than 50 neighbors in the set of points of all vehicular trips within 100 m radius as outliers. The entire vehicular trip is rejected once there are a considerable proportion of points in a vehicular trip (i.e., more than 20%) labeled as outliers. This method might also remove the trips taking the routes which are rarely used. But it would not impact our analysis as we place emphasis on the commonly used routes by each user. The pseudocode for deriving the vehicular trips from the raw LBS data is depicted in Algorithm 1.
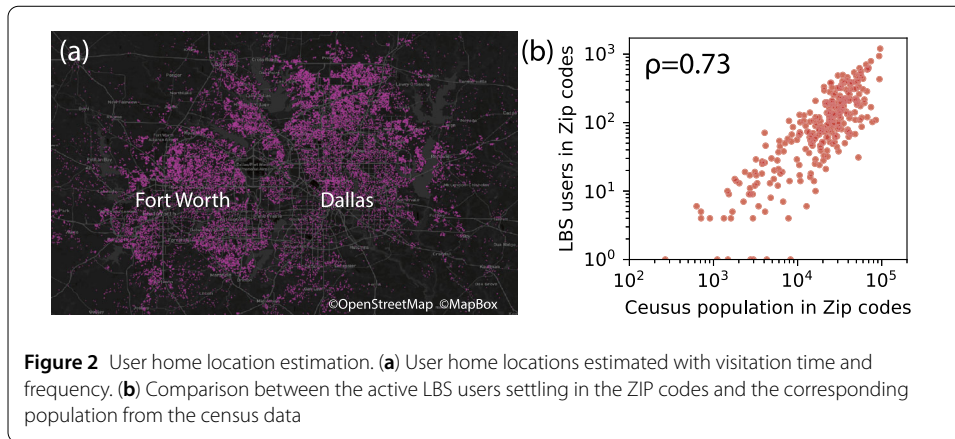

Algorithm 1 Vehicular Trips Deriving

1: *Step 1: User selection*
2: **for** *user* $\in$ *Users* **do**   ▷ Loops through each user in the raw data
3:       Count the timespan and the number of records
4:       **if** *timespan* < 60 days or # records < 1000, **then**
5:             Remove the user's records
6: *return* selected data
7: *Step 2: Vehicular trips detection*
8: **for** *user* $\in$ *Users* **do**   ▷ Loops through each user in the selected data
9:       **for** *item* $\in$ *records* **do**   ▷ Loop through each item in user's records
10:             Calculate the time difference between the current and next item, Δt
11:             **if** $\Delta t \geq 30$ min **then**
12:                   Partition the trace
13:       *User's trips are ready.*
14:       **for** *trip* $\in$ *Trips* **do**   ▷ Loop through each trip in user's trips
15:             Calculate the average travel speed, $\bar{v}$
16:             **if** $\bar{v} < 20$ km/h or $\bar{v} > 100$ km/h **then**
17:                   Remove the trip
18:             Update *tripId*
19: *return* vehicular trips
20: *Step 3: Vehicular trips filtering*
21: **for** *user* $\in$ *Users* **do**   ▷ Loops through each user with vehicular trips
22:       **for** *trip* $\in$ *Trips* **do**   ▷ Loop through each trip in user's trips
23:             Number of removed points, $N_{outlier} = 0$
24:             **for** *point* $\in$ *trip* **do**
25:                   Count neighboring points within 100 m in the data of all selected users,
         $N_{neighbor}$
26:                   **if** $N_{neighbor} < 50$ **then**
27:                         $N_{outlier}$ += 1
28:             **if** $N_{outlier}/length(trip) \geq 0.2$ **then**
29:                   Remove the trip
30:             Update *tripId*
31: *return* selected vehicular trips

3.2 Collective travel demand estimation and validation

Similar to travel behavior analysis using CDR data [20, 38, 39], we first detect the possible home locations of each active LBS user. To this end, we collect the stay locations for each user from the origins and destinations of all vehicular trips. Each stay location is associated with the departure or arrival time. As the users usually depart from home in the morning and arrive home in the end of the day, we select all departure locations between 5:00 a.m. and 10:00 a.m. and the arrival locations after 5:00 p.m. every day to compose the user's home candidate pool. If more than 30 locations are found, we then cluster the candidate locations using DBSCAN, setting the spatial threshold as 300 m, considering the users may park their vehicles around the significant places. We define the centroid of the largest cluster as a home place if the fraction of points in this cluster is larger than 40%.

Figure 2 User home location estimation. (**a**) User home locations estimated with visitation time and frequency. (**b**) Comparison between the active LBS users settling in the ZIP codes and the corresponding population from the census data

In Fig. 2(a), we show the detected home locations of the active LBS users. The accuracy of home detection is always challenging due to the lack of ground truth. Vanhoof et al. used the census data to validate the home location detection methods [40]. However, the validation can not be very reliable even at collective level because of the heterogeneous distribution of active LBS users in space. Note that we use the LBS users' home locations at ZIP code level to expand the users' travel demand, suggesting that we do not need to identify the home location within a few meters. We compare the active LBS users settling in each ZIP code versus its population from the U.S. census data [41], and find positive Pearson correlation ($\rho = 0.73$), as shown in Fig. 2(b). However, there do exist some ZIP codes with large populations but very small numbers of active LBS users, indicating the difficulty of user expansion.

Besides, we aggregate the derived vehicular trips by an hour to achieve the hourly flow during one week, as shown in Fig. 3(a), displaying morning and evening peaks on weekdays and midday peaks on weekends. We next inspect the vehicle trips from LBS data by comparing the origin-destination (OD) flow with the travel survey conducted by the North Central Texas Council of Governments (NCTCOG) in 2014 [42]. As the LBS data are collected from a fraction of the population, like the CDR data [20, 39], we define the expansion factors in each ZIP code as the ratio of the population from 2016 U.S. census data to the number of LBS users living in the same region. The distribution of expansion factors is presented in Fig. 3(b), and the 1st, 2nd, and 3rd quartiles of the expansion factors are 136.6, 217.2, and 356.8, respectively. Note that population synthesis is one more advanced way to expand the active users to population level than our expansion factor. The population synthesis expands the users with their demographic/socioeconomic information and sophisticated models [43]. The expansion factor of each Zip code is visualized in Fig. 3(c). The Zip codes in the urban area generally have smaller expansion factors than rural area. We then aggregate the vehicular trips during the morning peak hours (6:30-9:00 a.m.) at Zip code level and scale the flow with the expansion factors. In Fig. 3(d), we compare the values of vehicular travel demand for all OD pairs at ZIP code level between the expanded LBS flow and the NCTCOG survey in the morning peak hours, and find the linear fitting slop equals to 0.82 and $r^2 = 0.59$. Figure 3(e) illustrates the spatial distribution of vehicular travel flow above 0.01% of the total demand during the morning peak hours achieved from the expanded LBS data and NCTCOG data, respectively. Even we show the estimated travel demand is visually comparable to the NCTCOG survey data, the Pearson
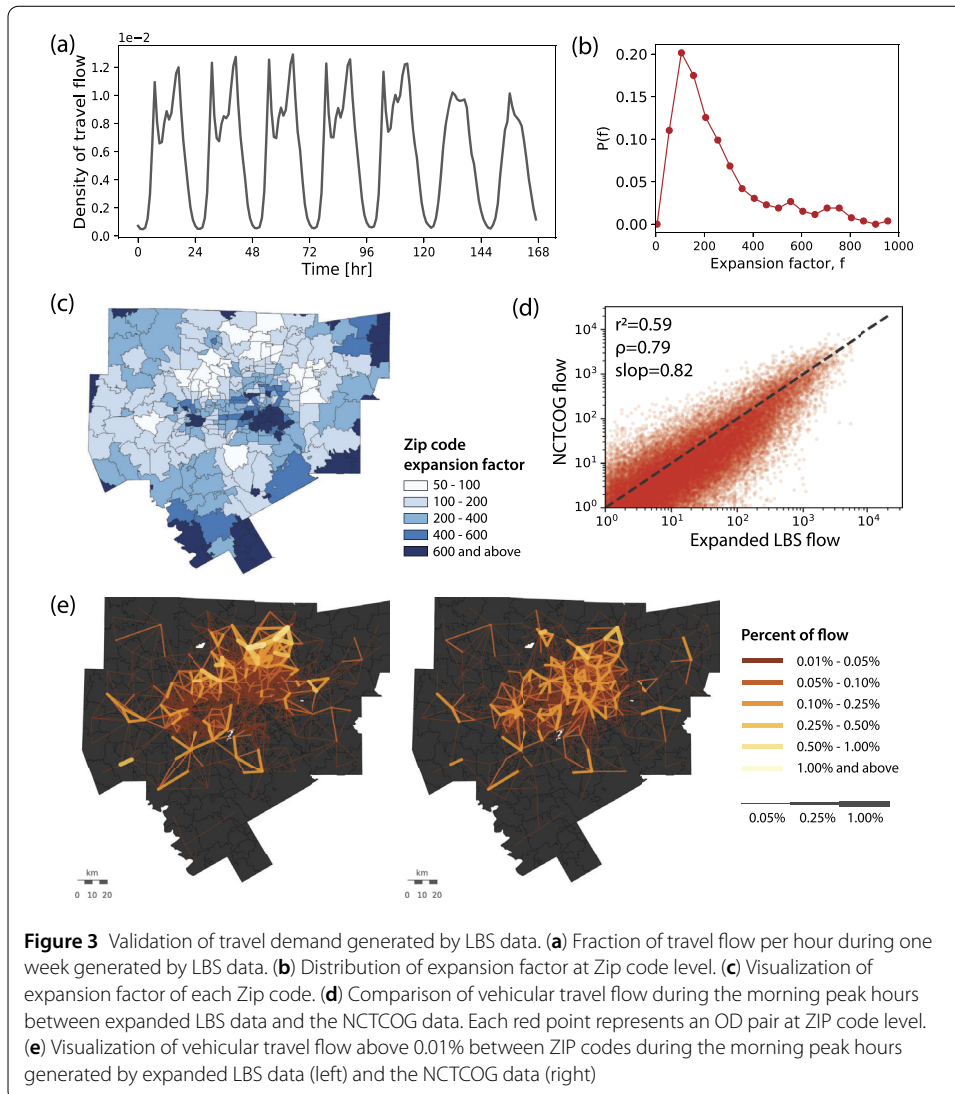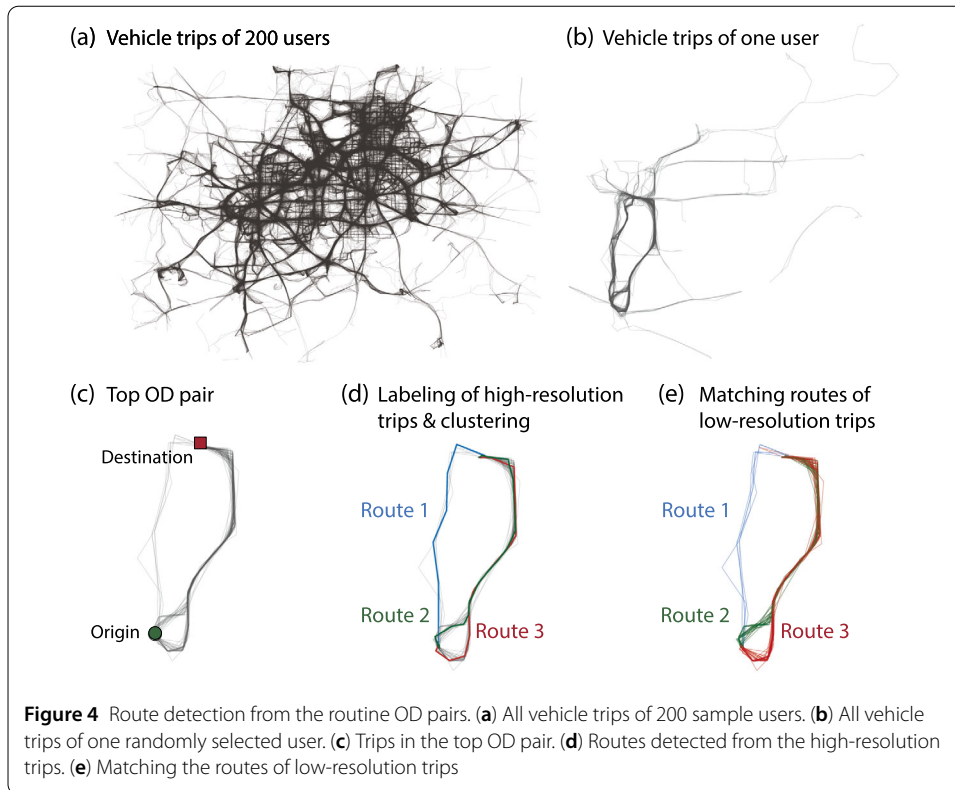
Figure 3 Validation of travel demand generated by LBS data. (**a**) Fraction of travel flow per hour during one week generated by LBS data. (**b**) Distribution of expansion factor at Zip code level. (**c**) Visualization of expansion factor of each Zip code. (**d**) Comparison of vehicular travel flow during the morning peak hours between expanded LBS data and the NCTCOG data. Each red point represents an OD pair at ZIP code level. (**e**) Visualization of vehicular travel flow above 0.01% between ZIP codes during the morning peak hours generated by expanded LBS data (left) and the NCTCOG data (right)

correlation only reaches 0.79. That can be caused by several reasons in this work, such as (i) we simply selected active users with the timespan and number of records in the raw LBS data, aiming at removing the temporary visitors in the DFW metroplex. However, we can not accurately identify residents from all LBS users with such simple rules; (ii) The distribution of active LBS users is different to the residents in space, as shown in Fig. 2(b) and the spatial distribution of expansion factors in Fig. 3(c); (iii) As we used simple rules to identify the vehicular trips, some non-vehicular trips are kept in our OD matrix; (iv) we used simple expansion factors to expand the travel demand of active LBS users to the population.

3.3 Route detection

The core challenge of deriving route choice behavior from LBS data is the varying sample interval of the records, as shown in Fig. 1(c). The varying sample interval in time leads to the heterogeneity of displacement between two consecutive records, ranging from several meters to kilometers. Such heterogeneity would cause the incorrect calculation of the similarity between two trips, and affect the clustering of trips. For instance, even when

Figure 4 Route detection from the routine OD pairs. (**a**) All vehicle trips of 200 sample users. (**b**) All vehicle trips of one randomly selected user. (**c**) Trips in the top OD pair. (**d**) Routes detected from the high-resolution trips. (**e**) Matching the routes of low-resolution trips

two low-resolution trips are taking the same route, the similarity between them would be low as the distance between the point pair would be large. One of the popular solutions is to map the points to the road network with map-matching and connect the distant consecutive points with the shortest path in the road network. However, it requires the road network and map-matching is computationally expensive for massive trajectory data [44].

To overcome this challenge, we design a simple yet efficient two-step procedure to find the routes: (i) selecting the high-resolution trips, in which the maximum distance gap between consecutive points is less than 1 km and detect the taken routes by trace clustering; (ii) matching the low-resolution trips to the high-resolution ones and finding the most likely taken routes. Figure 4(a) presents the extracted vehicular trips from the raw LBS records for 200 sample users. The layout of vehicular trips displays a good match with the road networks in the DFW metroplex. We select one user to illustrate the two-step procedure, as shown in Fig. 4(b). For understanding the route choice behavior, we decide to focus on the frequently visited places between which there are repeated number of trips. To this end, we cluster the origins and destinations of all trips using DBSCAN and label the centroids of clusters as significant places. For each active LBS user, we then select two unidirectional OD pairs for further route detection, the OD pairs with the largest and the second largest numbers of trips, as illustrated in Fig. 4(c). Note that the two selected OD pairs may not be reversed. After this step, we keep 1,194,154 trips (5.3% of all trips after trip segmentation) of 58,333 users (0.9% of all users in the raw LBS data) for routing behavior analysis.

Among the trips between a selected origin and destination pair, we first select the high-resolution trips to label the routes. This is because the inference is more reliable when the distance gaps between consecutive points are small. The high-resolution trips are grouped

to one or more clusters using a clustering method described in the following if there is more than one trip. The purpose of trip clustering is to group the trips which are taking the same route. There are two selection criteria for trip clustering, measurement of trip similarity and the number of clusters. Two of the most popular measurements are the longest common subsequence (LCSS) and dynamic time warping (DTW) [32, 45]. However, Atev et al. proposed a modified Hausdorff distance and confirmed that it could surpass both LCSS and DTW in trajectory clustering [46]. In fact, we find that the modified Hausdorff improves its robustness to the noise by rejecting a number of worst matches of points in the two trajectories. In this work, we adopt the same modified Hausdorff to calculate the distance between two high-resolution trips and DBSCAN to cluster them into one or more groups. Ideally, each cluster represents one route. For the sample user in Fig. 4(b), there are three routes detected on the high-resolution trips, differentiated by color in Fig. 4(d).

In the second step, we add the records of the low-resolution trips by aligning them with the high-resolution ones. Specifically, for each low-resolution trip, we first calculate the maximum distance between the point sets in it and the sets in each high-resolution trip. This distance indicates how far does this trip deviate from the high-resolution cluster and is used to decide if they belong to the same route. If the target low-resolution trip has a nearest high-resolution trip within a certain distance (e.g., 1 km), we identify its route the same as the high-resolution one. Otherwise, we remove this low-resolution trip as its route is uncertain. Figure 4(e) presents the final route detection results for the sample user. The detailed pseudocode for route detection is depicted in Algorithm 2.

4 Route choice behavior analysis

4.1 Distribution of number of routes

We first inspect the statistical distributions of the number of trips N_{trip} and the number of routes N_{route} in the selected top two OD pairs for all active users. Figure 5(a) and (b) present the distributions of N_{trip} and N_{route}, respectively. Log-normal distributions resemble the data in both cases, in agreement with Lima et al.'s findings [18]. The mean value of N_{trip} reaches 29.08, while the mean value of the N_{route} between these OD pairs is 1.56. From Fig. 5(b), we observe that among all active users in our LBS data, 51.35% of them only take one route to complete the top OD pairs; 37.5% of them take two routes and only 11.15% of them take more than 2 routes.

Next, we inspect the discrepancy of routing behavior during peak and off-peak hours. To this end, we split all trips in users' top OD pairs into four groups by their departure time, e.g., morning peak hours from 7:00 to 10:00 (AM), midday from 10:00 to 16:00 (MD), evening peak hours from 16:00 to 19:00 (PM) and the rest of the day (RD). We then count N_{route} of each user in these four time periods on weekdays and weekends, respectively. Figure 5(c) presents the fraction of active users taking different N_{route} during each time period on weekdays. The number of trips in each time period is presented in the inset. We observe the fractions of users taking 2 routes and above during AM and PM are apparently higher than the other two periods, suggesting that the users tend to take more routes during the peak hours to finish their trips more efficiently, e.g., in shorter travel time. As for the routing behavior on weekends shown in Fig. 5(d), the distribution of N_{route} changes little between time periods due to the traffic on weekends is not as congested as weekdays.

Algorithm 2 Route Detection
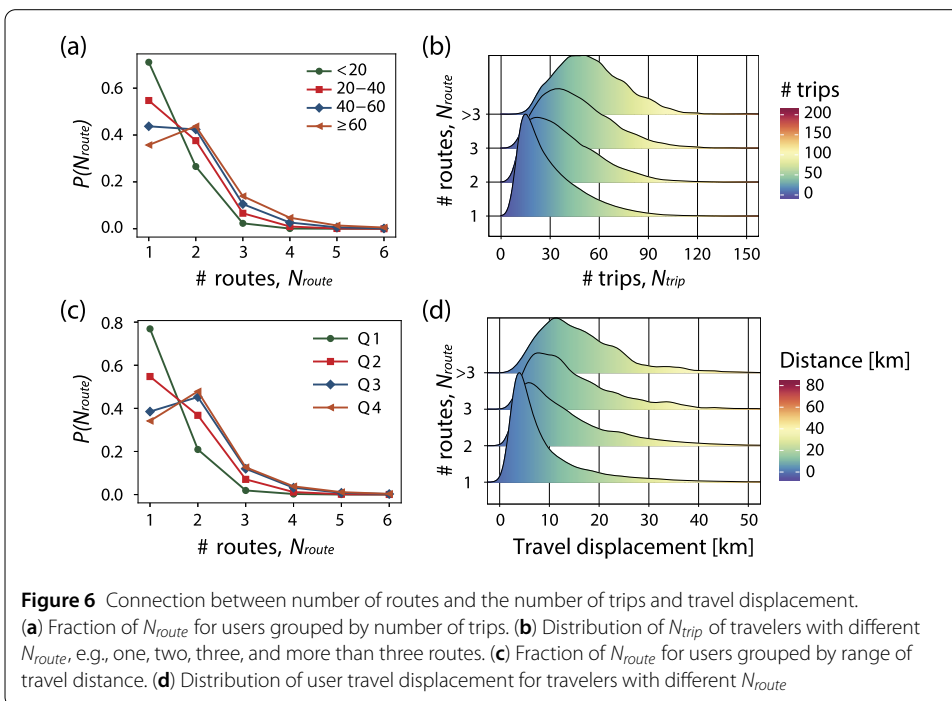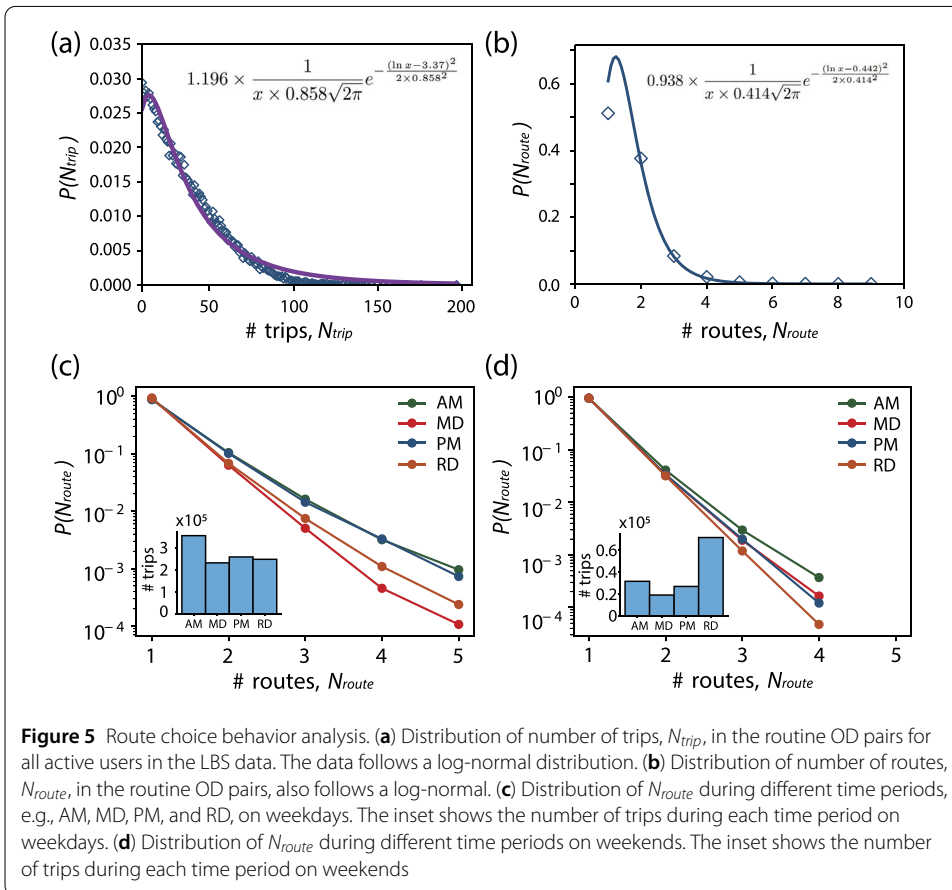

1: *Step 1: Top OD pair selection*

2: *1.1:* Find stay locations via DBSCAN clustering the origins and destinations

3: **if** # points in cluster ≥ 5, **then**

4:         Keep the centroid of cluster as a stay location

5: *1.2:* Count the number of trips between stay locations

6: *return* OD pairs with the largest and the second largest number of trips

7: *Step 2: Route detection on high-resolution trips*

8: *2.1:* Finding high-resolution trips

9: **for** *trip* $\in$ *TopODpair* **do**   ▷ Loops through each trip in the selected trips

10:        Calculate the distance gap between consecutive points, find the maximum gap g_{max}

11:        **if** $g_{max} \leq 1$ km **then**

12:               Label the trip as high resolution

13: *2.2:* Do trajectory clustering on all high-resolution trips using DBSCAN

14: *2.3:* Label each cluster as one route

15: *return* high-resolution trips with labeled routes

16: *Step 3: Route labeling for low-resolution trips*

17: **for** $trip_L \in$ low-resolution trips **do**   ▷ Loops through each trip in the low-resolution trips

18:        **for** $trip_H \in$ high-resolution trips **do**

19:               Calculate the maximum distance from points in $trip_L$ to $trip_H$

20:        Find the nearest $trip_{H,nearest}$ and the distance $D_{H,nearest}$

21:        **if** $D_{H,nearest} \leq 1$ km **then**

22:               Give the route label of $trip_{H,nearest}$ to $trip_L$

23:        **else**

24:               Route taken by $trip_L$ is uncertain, remove it

25: *return* trips with labeled routes

4.2 Route choice behavior of different groups of travelers

For comprehensive understanding of the discrepancy of route choice behavior among the travelers, we group them by their travel frequencies and travel distances. According to the distribution of N_{trip} presented in Fig. 5(a), we split the travelers into four groups according to the following rules, $N_{trip} < 20$, $20 \leq N_{trip} < 40$, $40 \leq N_{trip} < 60$, and $N_{trip} \geq 60$. Figure 6(a) presents the distribution of N_{route} per group. We notice that frequent travelers tend to explore more routes than non-frequent travelers, most likely because the frequent travelers know the traffic congestion better and are more confident to find efficient alternatives. The phenomenon also can be confirmed in Fig. 6(b), which presents the distribution of N_{trip} of users taking different number of routes in their routine OD pairs. The median value of N_{trip} of travelers with more than three routes is around 60, while the median value of travelers sticking on one route is less than 30.

The distance between origin and destination could be one of the factors that affect the number of routes selected as the distance determines the number of candidate routes in a given road network. We then compare the routing behavior of travelers with different ranges of travel distance. The travelers are grouped into Q1 to Q4 by the 25th, 50th, 75th

Figure 5 Route choice behavior analysis. (**a**) Distribution of number of trips, N_{trip}, in the routine OD pairs for all active users in the LBS data. The data follows a log-normal distribution. (**b**) Distribution of number of routes, N_{route}, in the routine OD pairs, also follows a log-normal. (**c**) Distribution of N_{route} during different time periods, e.g., AM, MD, PM, and RD, on weekdays. The inset shows the number of trips during each time period on weekdays. (**d**) Distribution of N_{route} during different time periods on weekends. The inset shows the number of trips during each time period on weekends

Figure 6 Connection between number of routes and the number of trips and travel displacement. (**a**) Fraction of N_{route} for users grouped by number of trips. (**b**) Distribution of N_{trip} of travelers with different N_{route}, e.g., one, two, three, and more than three routes. (**c**) Fraction of N_{route} for users grouped by range of travel distance. (**d**) Distribution of user travel displacement for travelers with different N_{route}

percentiles of their travel displacements. Figure 6(c) depicts the distribution of N_{route} of each group, and Fig. 6(d) depicts the distribution of travel displacements of travelers with different N_{route}. As expected, we can observe that most of the users with short trips in Q1 stick on only one route. From Fig. 6(d), we can see the peak of the distribution is around 5 km for the users who only take one route; while the peak is around 10 km for the users who take more than 3 routes. These observations indicate that more routes are likely been taken if the users make longer trips between two significant places. It can be explained from the perspective of network. In a dense road network, the larger the distance between two nodes is, the more alternative routes with similar cost the travelers can choose.

4.3 Route choice behavior in traffic congestion

Traffic congestion is a major consideration driving travelers to find alternatives, especially during peak hours. Here, we investigate the relation between travel time and the number of routes in the routine OD pairs of active travelers. For each user, we calculate the travel time index (TTI) of all trips in travelers' top OD pairs to assess the additional travel time caused by congestion. Given a number of trips in an OD pair, TTI is defined as the ratio of the average travel time to the free flow travel time,

$$TTI = T_{avg}/T_{free}, \tag{1}$$

where T_{avg} refers to the average travel time of all trips made by one user; T_{free} refers to the free flow travel time from her origin to destination, approximated by the minimum travel time among all trips in the OD pair. The larger the TTI is, the more traffic congestion the user met during the routine journey. The introducing of TTI enables us to compare the travel delay of OD pairs even they have various travel distances. Figure 7(a) and (b) illustrate the TTI of the travelers with different N_{route} during the AM (7:00-10:00) and PM (16:00-19:00) peak hours on weekdays, respectively. It is noticeable that travelers taking more routes tend to have lower TTI. The average TTI values are also illustrated in Table 1. To reduce the impact of the extreme small or large values of TTI, we also present the mean value of the TTI falls into the 95% confidence interval and the standard deviation (STD) in Table 1. We can conclude that travelers with flexible route choice behavior can lower their travel time by avoiding traffic congestion in the primary routes. The insets of Fig. 7(a) and (b) present the distribution of TTI for all travelers, showing the average TTI is nearly 2.0 during the peak hours. This reveals that, because of the congestion, the travelers in DFW metroplex spent nearly double free flow travel time to complete their journeys during rush hours.

Beyond the additional traffic time caused by congestion, the reliability of travel time is also significant to many travelers, especially when they need to arrive at the destination on time. Reliability has been considered as a key performance measure by transportation planners and decision-makers. We introduce the buffer index (BI) to assess the travel time reliability of all trips in the traveler's top OD pair [47]. The BI represents the extra buffer time that the traveler should add to the average travel time when planning trips to ensure on-time arrival. Here we define BI as the relative gap between the 85th percentile travel time and the average travel time of an OD pair,

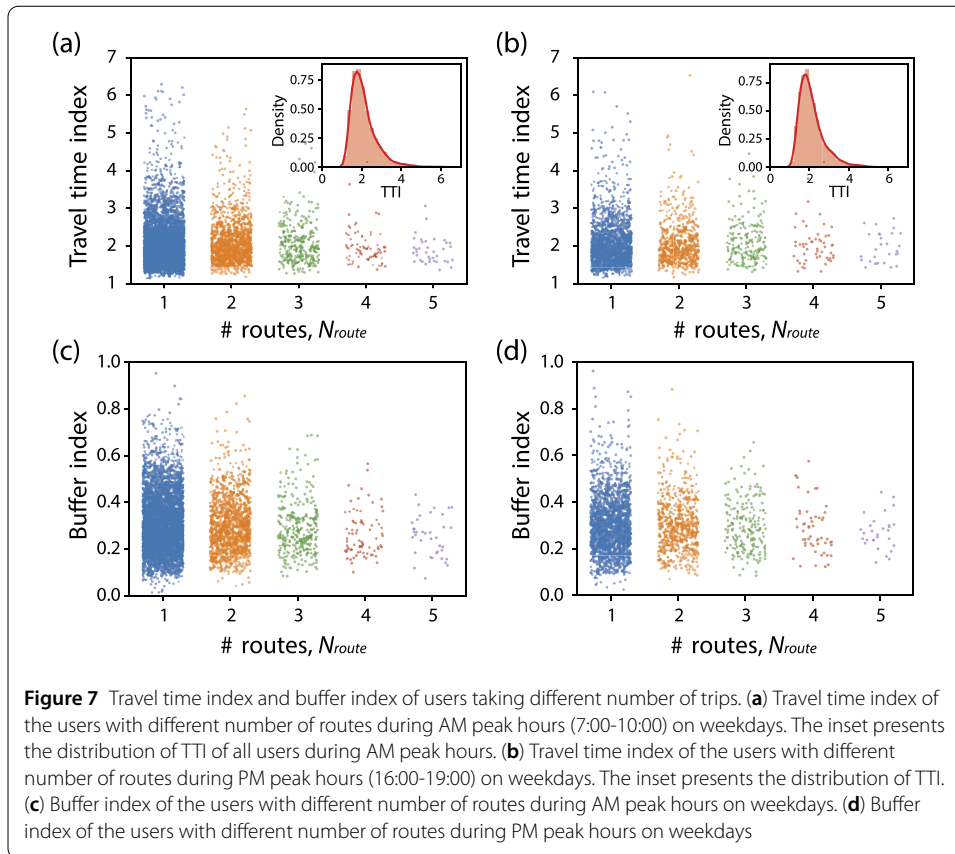$$BI = (T_{85th} - T_{avg})/T_{avg} \times 100\%. \tag{2}$$

Figure 7 Travel time index and buffer index of users taking different number of trips. (**a**) Travel time index of the users with different number of routes during AM peak hours (7:00-10:00) on weekdays. The inset presents the distribution of TTI of all users during AM peak hours. (**b**) Travel time index of the users with different number of routes during PM peak hours (16:00-19:00) on weekdays. The inset presents the distribution of TTI. (**c**) Buffer index of the users with different number of routes during AM peak hours on weekdays. (**d**) Buffer index of the users with different number of routes during PM peak hours on weekdays

Table 1 Mean values and STD of travel time index and buffer index of users taking different number of trips. The **TTI*** and **BI*** indicate the mean values are calculated in the 95% confidence intervals

# routes	1	2	3	4	5
average TTI (AM peak)	2.110	2.139	2.095	1.991	1.878
average **TTI*** (AM peak)	2.042	2.072	2.050	1.940	1.847
STD of TTI (AM peak)	0.669	0.654	0.556	0.468	0.341
average TTI (PM peak)	2.077	2.115	2.104	2.059	1.952
average **TTI*** (PM peak)	2.009	2.044	2.060	1.998	1.936
STD of TTI (PM peak)	0.654	0.696	0.524	0.552	0.389
average BI (AM peak)	31.46%	30.50%	28.97%	26.67%	25.49%
average **BI*** (AM peak)	30.89%	29.91%	28.23%	26.03%	25.50%
STD of BI (AM peak)	12.84%	12.04%	11.52%	9.57%	8.50%
average BI (PM peak)	30.63%	30.41%	29.83%	28.84%	27.12%
average **BI*** (PM peak)	29.99%	29.72%	29.33%	28.33%	26.82%
STD of BI (PM peak)	12.52%	11.87%	11.00%	10.76%	7.22%

The BI is expressed as a percentage and its value increases as reliability gets worse. Figure 7(c) and (d) present the BI of the travelers with different N_{route} during the AM and PM peak hours on weekdays, respectively. The average BI values, the average BI in the 95% confidence interval, and the STD of BI are illustrated in Table 1. We notice that the average BI decreases along with the increase of N_{route}, suggesting that the travelers are changing their routes considering the real-time traffic to increase the reliability of travel time.

5 Conclusion and outlook

Understanding the route choice behavior is an essential task for not only modeling human mobility in transportation networks but also route management to relieve traffic congestion. In this paper, we presented a data analysis framework for understanding route choice behavior with massive LBS data. Steps include, user selection, trip clustering, route detection, and behavior analysis. We analyzed the six-month LBS data in the Dallas-Fort Worth metroplex, and selected the trips between the most frequent origin-destination pair of each user for understanding routing behavior. We found that the distribution of the number of routes can be modeled by a log-normal distribution. We also inspected the relation between the number of routes and the travel displacement and found that travelers with longer travel distances tend to select more routes to shorten their travel time. We also confirmed that travelers take more routes during peak hours than off-peak hours, and those individuals that explore more routes reduce their impact of congestion and increase their reliability of travel times. The proposed framework makes LBS data useful to evaluate the route choice behavior of different groups of travelers and their reaction to traffic congestion. As future applications, this could be implemented to evaluate traffic regulation strategies, such as the congestion charges.

Moreover, there are still some directions for further study. For instance, (i) for comprehensively understanding people's emphasis on different travel costs (i.e., travel time, routing distance, etc.), we need to further integrate these factors per route per user. To that end, we need to estimate the traffic states in the entire road network through mapmatching; (ii) this work presented a case study in a region without congestion pricing. The relation between socio-economic characteristics and routing behavior merits more attention in cities with traffic interventions.

Authors' contributions
YX and MCG conceived the research and designed the analyses. YX and RDC processed and analyzed the data. YX and MCG performed the results analyses and wrote the paper. MCG provided general advice and supervised the research. All authors read and approved the final manuscript.

Author details
[1]MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, 200240, Shanghai, China. [2]Department of Civil & Environmental Engineering, MIT, 02139, Cambridge, MA, USA. [3]Department of City and Regional Planning, University of California, Berkeley, 94720, Berkeley, CA, USA. [4]Energy Analysis & Environmental Impacts Division, Lawrence Berkeley National Laboratory, 94720, Berkeley, CA, USA. [5]Centre for Advanced Spatial Analysis, University College London, London, WC1E 6BT, UK. [6]Department of Computer Science, University of Exeter, Exeter, EX4 4NR, UK. [7]Department of Civil and Environmental Engineering, University of California, Berkeley, 94720, Berkeley, CA, USA.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Xu Y, Olmos LE, Abbar S, González MC (2020) Deconstructing laws of accessibility and facility distribution in cities. Sci Adv 6(37):4112
2. Weisbrod G, Vary D, Treyz G (2003) Measuring economic costs of urban traffic congestion to business. Transp Res Rec 1839:98–106
3. Jiang B, Liang S, Peng Z-R, Cong H, Levy M, Cheng Q, Wang T, Remais JV (2017) Transport and public health in China: the road to a healthy future. Lancet 390(10104):1781–1791
4. Xu Y, Jiang S, Li R, Zhang J, Zhao J, Abbar S, González MC (2019) Unraveling environmental justice in ambient $PM_{2.5}$ exposure in Beijing: a big data approach. Comput Environ Urban Syst 75:12–21
5. Hu T-Y, Mahmassani HS (1997) Day-to-day evolution of network flows under real-time information and reactive signal control. Transp Res, Part C, Emerg Technol 5(1):51–69
6. Kelly FJ, Zhu T (2016) Transport solutions for cleaner air. Science 352(6288):934–936
7. Xu Y, González MC (2017) Collective benefits in traffic during mega events via the use of information technologies. J R Soc Interface 14(129):20161041
8. Çolak S, Lima A, González MC (2016) Understanding congested travel in urban areas. Nat Commun 7(1):1–8
9. Prud'homme R, Bocarejo JP (2005) The London congestion charge: a tentative economic appraisal. Transp Policy 12(3):279–287
10. Wu C, Bayen AM, Mehta A (2018) Stabilizing traffic with autonomous vehicles. In: 2018 IEEE international conference on robotics and automation (ICRA). IEEE, pp 1–7
11. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779
12. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6:8166
13. Jiang S, Yang Y, Gupta S, Veneziano D, Athavale S, González MC (2016) The TimeGeo modeling framework for urban mobility without travel surveys. Proc Natl Acad Sci USA 113(37):5370–5378
14. Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions. In: Handbook of transportation science. Springer, Boston, pp 5–33
15. Prato CG (2009) Route choice modeling: past, present and future research directions. J Choice Model 2(1):65–100
16. Di X, Liu HX (2016) Boundedly rational route choice behavior: a review of models and methodologies. Transp Res, Part B, Methodol 85:142–179
17. Zhu S, Levinson D (2015) Do people use the shortest path? An empirical test of wardrop's first principle. PLoS ONE 10(8):0134322
18. Lima A, Stanojevic R, Papagiannaki D, Rodriguez P, González MC (2016) Understanding individual routing behaviour. J R Soc Interface 13(116):20160021
19. Di Clemente R, Luengo-Oroz M, Travizano M, Xu S, Vaitla B, González MC (2018) Sequences of purchases in credit card data reveal lifestyles in urban populations. Nat Commun 9(1):1–8
20. Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, González MC (2015) The path most traveled: travel demand estimation using big data resources. Transp Res, Part C, Emerg Technol 58:162–177
21. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10(84):20130246
22. Xu Y, Çolak S, Kara EC, Moura SJ, González MC (2018) Planning for electric vehicle needs by coupling charging profiles with urban mobility. Nat Energy 3:484–493
23. Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS (2017) Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp Res, Part C, Emerg Technol 75:197–211
24. Scherrer L, Tomko M, Ranacher P, Weibel R (2018) Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth. EPJ Data Sci 7(1):19
25. Dong X, Morales AJ, Jahani E, Moro E, Lepri B, Bozkaya B, Sarraute C, Bar-Yam Y, Pentland A (2019) Segregated interactions in urban and online spaces. arXiv preprint. arXiv:1911.04027
26. Liao Y, Yeh S, Jeuken GS (2019) From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. EPJ Data Sci 8(1):34
27. McNeill G, Bright J, Hale SA (2017) Estimating local commuting patterns from geolocated Twitter data. EPJ Data Sci 6(1):24
28. Aleta A, Piontti APY, Ajelli M, Litvinova M et al Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scen-arios of the covid-19 epidemic. Technical report
29. Klein B, Privitera F, Lake B, Kraemer MU, Brownstein JS, Lazer D, Eliassi-Rad T et al (2020) Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak
30. Kwan M-P (2004) GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. Geogr Ann, Ser B, Hum Geogr 86(4):267–280
31. Ratti C, Frenchman D, Pulselli RM, Williams S (2006) Mobile landscapes: using location data from cell phones for urban analysis. Environ Plan B, Plan Des 33(5):727–748
32. Zheng Y (2015) Trajectory data mining: an overview. ACM Trans Intell Syst Technol 6(3):29
33. Miller HJ, Goodchild MF (2015) Data-driven geography. GeoJournal 80(4):449–461
34. Silva TH, Viana AC, Benevenuto F, Villas L, Salles J, Loureiro A, Quercia D (2019) Urban computing leveraging location-based social network data: a survey. ACM Comput Surv 52(1):17
35. Cuebiq Offline Intelligence Measurement https://www.cuebiq.com. [Online; accessed September-2019] (2019)
36. Xiao G, Juan Z, Zhang C (2015) Travel mode detection based on GPS track data and Bayesian networks. Comput Environ Urban Syst 54:14–22

37. Dabiri S, Heaslip K (2018) Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp Res, Part C, Emerg Technol 86:360–371
38. Jiang S, Fiore GA, Yang Y, Ferreira J Jr, Frazzoli E, González MC (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, New York, p 2
39. Çolak S, Alexander LP, Alvim BG, Mehndiratta SR, González MC (2015) Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transp Res Rec 2526:126–135
40. Vanhoof M, Reis F, Ploetz T, Smoreda Z (2018) Assessing the quality of home detection from mobile phone data for official statistics. J Off Stat 34(4):935–960
41. U.S. Census Bureau https://www.census.gov/. [Online; accessed September-2018] (2016)
42. The North Central Texas Council of Governments https://www.nctcog.org/. [Online; accessed September-2018] (2014)
43. Sun L, Erath A (2015) A Bayesian network approach for population synthesis. Transp Res, Part C, Emerg Technol 61:49–62
44. Chen BY, Yuan H, Li Q, Lam WH, Shaw S-L, Yan K (2014) Map-matching algorithm for large-scale low-frequency floating car data. Int J Geogr Inf Sci 28(1):22–38
45. Kim J, Mahmassani HS (2015) Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. Transp Res, Part C, Emerg Technol 59:375–390
46. Atev S, Miller G, Papanikolopoulos NP (2010) Clustering of vehicle trajectories. IEEE Trans Intell Transp Syst 11(3):647–657
47. FHWA https://ops.fhwa.dot.gov/publications/tt_reliability/TTR_Report.htm. [Online; accessed September-2019] (2019)