Transportation Research Part C 118 (2020) 102709

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Estimating MFDs, trip lengths and path flow distributions in a multi-region setting using mobile phone data

Mahendra Paipuri[a,*], Yanyan Xu[b], Marta C. González[b], Ludovic Leclercq[a]

[a] Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT, Lyon, France
[b] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Keywords:
MFD
Mobile phone data
Penetration rates
Map matching
Macro-paths
Dynamic trip length
Path flow distribution
User Equilibrium gap

ABSTRACT

The present work proposes a global framework to estimate *all* MFD model parameters using mobile phone data. The three major components that are estimated in the present context are MFD shapes, regional trip lengths and path flow distribution. A trip enrichment scheme based on the map matching process is proposed for the trips that have sparser records. Time dependent penetration rates are estimated by fusing the OD matrix and the Loop Detector Data (LDD). Two different types of penetration rates of vehicles are proposed based on the OD flow and the trips starting within an origin, respectively. The estimated MFDs based on two types of penetration rates are stable with very low scatter. In the following step, macro-paths and their corresponding trip lengths are estimated. This work is the first to present empirical evidences of the dynamic evolution of mean trip lengths over the day, which is very difficult to capture with other types of data sources. The last component is the time dependent path flow distributions between the different macro-paths for a given OD pair. The manuscript is concluded by presenting the time evolution of the User Equilibrium (UE) gap for different macroscopic OD pairs. It is noticed that UE principle holds true most of the time, except for OD pairs that have macro-paths transversing through congested reservoirs, especially during peak hours.

1. Introduction

The Macroscopic Fundamental Diagram (MFD) relates average density to average flow in the urban network (Mahmassani et al., 1984; Daganzo, 2007) and has evolved as a promising tool for urban traffic management. The empirical existence of MFD was first reported by Geroliminis and Daganzo (2008) for the city of Yokohama, Japan under certain homogeneity assumptions. Since then, several applications like traffic state estimation (Knoop and Hoogendoorn, 2014; Yildirimoglu and Geroliminis, 2014; Kavianipour et al., 2019), perimeter control (Haddad and Mirkin, 2017; Ampountolas et al., 2017; Sirmatel and Geroliminis, 2018; Mohajerpoor et al., 2019), congestion pricing (Gu et al., 2018), route guidance (Yildirimoglu et al., 2015) and cruising for parking (Cao and Menendez, 2015; Leclercq et al., 2017), *etc.* have been proposed based on the MFD approach.

Accurate estimates of the MFD, trip length distributions and path flow coefficients are crucial for predicting accurate predictions from MFD models. Most of the applications founded on the MFD assume a well-defined MFD relation for the urban network under consideration. However, in reality, the estimation of the MFD for urban networks is far from trivial. There are two main types of data sources namely, Loop Detector Data (LDD) and Floating Car Data (FCD) to estimate the MFD. Most of the empirical MFDs proposed in

* Corresponding author.
E-mail addresses: mahendra.paipuri@univ-eiffel.fr (M. Paipuri), yanyanxu@berkeley.edu (Y. Xu), martag@berkeley.edu (M.C. González), ludovic.leclercq@univ-eiffel.fr (L. Leclercq).

https://doi.org/10.1016/j.trc.2020.102709
Received 20 January 2020; Received in revised form 19 June 2020; Accepted 23 June 2020

the literature are based on LDD or FCD or a combination of both data sources (Ambühl et al., 2017). Wang et al. (2015) and Ampountolas and Kouvelas (2015) used LDD to estimate empirical MFDs for the urban networks of Sendai, China and Chania, Greece, respectively. The main limitation of the data derived from LDDs is the placement and distribution of loop-detectors. Buisson and Ladier (2009) demonstrated that the slope of the MFD depends on the distance of loop-detectors from downstream traffic signals. Although a methodological framework was proposed by Leclercq et al. (2014) to compute the average density on the link based on the placement of loop-detectors, network coverage remains a major limitation when calibrating an accurate MFD (Courbon and Leclercq, 2011; Ambühl and Menendez, 2016). More recently, Shim et al. (2019) studied the bifurcations in empirical an MFD that was estimated by roadside detectors while Alonso et al. (2019) analyzed the shape of empirical MFD for urban corridors using LDD. On the other hand, FCD is more attractive than its counterpart, as it provides vehicular trajectory data. Typically, FCD is provided by taxis either by Global Positioning System (GPS) or mobile phones. FCD was used to estimate and empirical MFD (Geroliminis and Daganzo, 2008; Bazzani et al., 2011; Tsubota et al., 2014), traffic monitoring (Herrera et al., 2010) and travel time estimation (Jie et al., 2011). Beibei et al. (2018) and Ambühl and Menendez (2016) estimated an empirical MFD by merging LCD and FCD data. Shoufeng et al. (2013) used the combination of GPS data and visually counted traffic to estimate the MFD for the Central Business District (CBD) of the city of Changsha in China. As stated above most of the works that employ FCD use only GPS data from taxis, as the GPS data of private cars are not readily available. Hence, the penetration rate of taxis is a potential limiting factor for estimating robust MFD from FCD (Du et al., 2016). Knoop et al. (2018) used large scale FCD to estimate the empirical MFD of Amsterdam by assuming a constant penetration rate. Ji and Geroliminis (2012, in press) discussed the importance of computing accurate penetration rates for MFD estimation using the GPS data of taxis. Recently, Huang et al. (2019) used the GPS data from taxis, private cars and public buses to estimate a 3D-MFD for the city of Shenzhen, China.

Another key ingredient of the MFD-based modeling framework is the set of macro-paths and their corresponding trip length distributions. It is not trivial to estimate either macro-paths or their trip lengths using LDD without any other equipment. On the other hand, FCD from taxis can be processed to obtain a distribution of trip lengths. However, FCD are generally sparse and fail to capture the repetitive trips frequently made by users/residents. Most of the works proposed in the context of the MFD-based framework assume a constant trip length inside the reservoir. However, it was concluded that using a single mean trip length inside the reservoir might introduce a significant error in traffic dynamics (Yildirimoglu and Geroliminis, 2014; Kouvelas et al., 2017) deduced. The importance of estimating accurate and reliable trip lengths in the context of MFD-based simulation was discussed in-detail in Batista et al. (2019, 2020). However, due to the lack of empirical data, the authors of the mentioned built a virtual set of shortest path trips by randomly sampling the origins and destinations in the network. Similarly, it is difficult to observe path flow distributions along each macro-path with existing LDD or FCD data sources. Yildirimoglu et al. (2015) used Dynamic Traffic Assignment (DTA) principle to estimate the path flow distribution in the context of route guidance. Most often DTA problems assume User Equilibrium (UE), Stochastic User Equilibrium (SUE) or Bounded Rational User Equilibrium (BRUE) conditions Batista et al. (2019). However, recent studies (Mariotte et al., 2020) have shown that UE route choice discipline may not be valid for macro-paths and large-scale networks.

As mobile phone data is increasingly readily available, they provide an attractive alternative to traditional data sources. Although numerous works have addressed the question of empirical MFD calibration, only a few of them specifically targeted the use of mobile phone data. This type of data has specific challenging characteristics like heterogeneous time-resolution and variable penetration rates in space and time. These issues are appropriately addressed in the framework proposed. The scope of the current work goes beyond the estimation of MFD shape only by proposing a unified framework to estimate *all* the required MFD model parameters, *i.e.*, trip lengths and path flow distributions from Location Based Service (LBS) data. This type of data is generated by the smartphone Apps, which share their location data actively with the App developer. The positioning of the user device is provided by either GPS or Wi-Fi. Recently, LBS data have been used to propose frameworks for data driven metrics like Origin–Destination (OD) matrix estimation (Jin et al., 2014), and travel route identification (Hsieh et al., 2015), *etc*. One obvious advantage of LBS over FCD data is that the former ensure wide coverage of the population in the urban network and therefore, high penetration rates are observed. As already stated, LBS data have a highly variable sampling interval, which can vary from a few seconds to a few minutes. A larger sampling interval implies a larger sampling distance and therefore inaccuracies in the distance traveled are inevitable in the case of urban networks. This work proposes a method to enhance the data with large sampling intervals by using map-matching techniques. In parallel, it also introduces a framework to estimate time dependent penetration rates. The secondary contribution of this work is the empirical estimation of trip length distributions. A static analysis is proposed, which yields the major macro-paths (or regional paths) along with their trip length distributions for a given OD pair. Yildirimoglu and Geroliminis (2014) showed that it is important to consider the dynamic variation of the trip lengths inside a region using micro-simulation studies. However, the empirical estimation of this so-called dynamic mean trip length is still an open question. Thus, dynamic analysis is also presented to study the variations in mean trip lengths during on-and off-peak hours. It is concluded that there exists a well-defined correlation between mean trip lengths and mean speeds at the regional level using the present data. The final part of the manuscript discusses the estimation of path flow distribution and UE gaps in the network. This gives valuable information about how far the network is from the UE state and it can help to better review equilibrium models at this scale. In addition, path flow coefficients and gap values can be used to validate the DTA models and calibrate the DTA parameters.

Overall, the contributions of the present work can be summarized as follows:

- A computationally efficient trip enrichment scheme is proposed to map-match the data and increase sampling frequency. One of the major drawbacks of LBS data is the variable sampling interval thus the technique proposed enhances the spatial and temporal resolution of trajectories. In addition, the method preserves the original pattern or structure of the underlying trip.
- Most works assume *a priori* a known constant penetration rate in the estimation of mean flow and density. This work proposes

time-dependent penetration rates at two different aggregation levels, namely OD-specific and origin-specific. Using empirical data it is shown that penetration rates are time-dependent and that assuming a time-invariant constant penetration rate significantly influences the shape of the estimated MFD.

- The importance of estimating the accurate distribution of trip lengths in the MFD-based framework has already been extensively discussed in the literature. The present work is the first to compute empirical trip length distributions using real data. Furthermore, it provides the first empirical evidence of the variation of mean trip length with time (dynamic mean trip length), which was previously shown only with simulation studies. Besides, this work is also first to propose a well-defined correlation between the trip detour ratio and mean speeds that can be used in MFD-based simulations to adjust the trip lengths with the network traffic states.
- Another important component in the MFD-based modeling framework are path flow coefficients. It is possible to observe path flow coefficients only at the regional level and they are often computed using DTA assuming UE conditions. The proposed framework estimates the empirical path flow distributions and concludes that they are time-dependent. In addition, UE gaps are estimated empirically, and can be used to validate the UE hypothesis in at the regional path level.

The paper is organized as follows: in Section 2 the details of data and data processing techniques are presented; then, in Section 3, the trip enhancement method used to improve the trajectory data for trips with sparse data points is discussed. Afterwards, a description of the computation of macroscopic variables and penetration rates is given in Section 4 followed in Section 5 by a presentation of the estimated MFDs and discussion on the penetration rates. In Section 6, the static and dynamic analysis of trip lengths are presented and, finally, Sections 7 and 8 consist of a discussion on the estimation of path flow distributions and UE gaps, respectively, based on the data.

2. Description and processing of data

2.1. Data details

The data contain the positioning of smartphone devices either by GPS or WiFi Positioning System (WPS) for the city of Dallas, Texas, in the United States, for a period of 2 months from March, 2017 to April, 2017. Dallas is one of the most populous cities in the US with an estimated 1.3 m inhabitants. In the present work, downtown Dallas and its neighboring suburbs as shown in Fig. 1a are considered to calibrate the MFD models. Fig. 1b shows the link level representation of the area under consideration, which extends across 160 km^2 and contains 18,386 nodes and 48,287 links. The length of the road network is 4800 km, which includes all types of roads. The raw data contain an anonymized user ID, timestamp, longitude, latitude and uncertainty regarding location. The data consists of 85,434 users and there are around 22 m records available for processing. Fig. 2a shows the visitation map for the area considered for a period of 14 days. It is clear from the heat map that the data are more concentrated in downtown Dallas and the major arterials surrounding downtown. LBS data are only generated when users interact with a location sharing application on smartphones. At the same time, different applications use location data at different frequencies. For instance, applications that use map-related services share their location more actively than others. Therefore, LBS data have different sampling intervals that range from a few seconds to a few minutes. A sampling interval is defined as the time difference between two consecutive records. Fig. 2b presents the probability distribution of the sampling intervals of the raw data for the month of March, 2017. Although a large fraction of users have sampling intervals around 100 s, it can be noticed from the distribution that there are peaks at 600 s and around 1000 s. A large sampling interval is a limitation while computing macroscopic variables like total traveled distance, trip length, *etc*. This issue

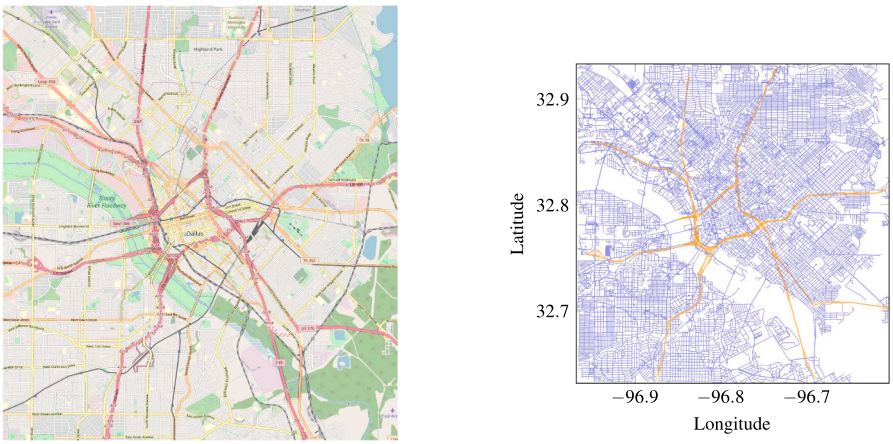

Fig. 1. Dallas network: map of the area and its link level description. (a) Map of the Dallas, TX ©OpenStreetMap 2019. (b) Link level representation of Dallas.

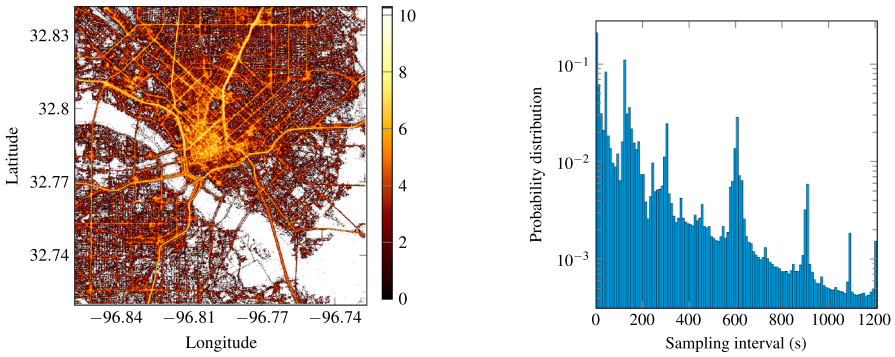

Fig. 2. Heat map and sampling rate distribution of the raw LBS data for selected days in March 2017. (a) Visitation map. The contour gives the log scale of the number of data records in the visitation map. (b) Distribution of sampling rates.

is addressed in the following sections by introducing a map matching scheme and by re-constructing the trajectory of the trip that has large sampling intervals.

2.2. Trip segmentation

Data selection and the segmentation of individual records into trips are discussed in detail in Xu et al. (2020) and a similar approach is used in this work. Data processing is done in various stages to extract useful and representative data. Firstly, records that have an uncertainty of more than 100 m are removed from the analysis. These LBS data contain different types of users like residents, tourists, people using highways, *etc*. Several users leave either very few records or stay for a very short period of time. In order to obtain representative trips from regular users of the urban network, it is necessary to filter the users with few records. Hence, only users with more than 1000 records that span across 30 days are considered in the present work. The resulting fraction of active users and selected records are 25% and 93%, respectively.

The following step consists in segmenting the records of each user into individual trips. This is done in two stages namely coarse segmentation and fine segmentation. As the names suggest, the first stage involves the coarse segmentation of selected records into trips. The second stage deals with the refinement of already segmented trips, either by splitting them into further trips or removing the records in the trips that do not comply with the user movement. In the coarse segmentation, the records are clustered into the trips based on the following assumptions. Firstly, the user starts a new trip if the time elapsed from the previous record is more than 30 min. Finally, only trips that have at least 5 records are selected to have more robust trip information. The second stage refines the already segmented trips in order to enhance trip quality. In this stage, the mean speed between consecutive records is monitored and those that have a mean speed of less than 4 km h^{-1} are removed from the trips. This method ensures that the records are clustered into representative individual trips and yields a total of 3.3 m trips.

2.3. Mode detection

The segmented trips can represent a variety of transport modes like subway, trams, buses, *etc*. Transportation mode detection is a crucial step in analyzing the GPS traces of mobile phone data. Many works that address the problem of mode detection can be found in the literature. They mainly fall within three different categories, namely machine learning algorithms (Reddy et al., 2008; Zhang et al., 2012; Montoya et al., 2015), probabilistic methods (Chen and Bierlaire, 2015) and decision criteria techniques (Mun et al., 2008; Gong et al., 2012). Although high detection accuracy was reported for most of the works, many challenges remain regarding research on mode detection (Feng and Timmermans, 2016; Nikolić and Bierlaire, 2017; Yang et al., 2018b). For instance, it is very difficult to characterize a slow-moving bicycle and a brisk walker. On the other hand, using additional Geographic Information System (GIS) tools can enhance the quality of results. However, this is achieved at a huge computational cost in order to process the data.

The analysis of the light-rail trips is done based on the decision criteria method proposed in Gong et al. (2012). There are a total of 4 light-rail lines that cover approximately 150 km with 64 stations. For each raw trip, the distances from the trip origin location to all the light-rail stations are estimated. Similarly, the distance from the destination to all the stations is also computed. If both origin and destination are within 100 m of two different light-rail stations, it is marked as a possible light-rail trip for further verification. Now, a distance matrix is computed between all the records of the selected trip to all the light-rail station locations. Using this estimated distance matrix, stations are identified where the distance between a particular station location to any record in the trip is less than 100 m. If the stations identified are within the origin and destination stations, the trip is marked as a light-rail trip. This simple method filters all the light-rail trips in the area considered. It can be noticed that this method yields a total number of 8652 of trips out of 3.3 m. This corresponds to a mere 0.2 % of total trips. A technique similar to that used to detect light-rail trips can be used for separating bus trips. However, current OpenStreetMap (OSM) GIS data lack information on all the bus stop locations in Dallas. In

addition, there are a total of 11000 bus stops in the Dallas area and thus the present decision based method will require a significant computation time to detect the trips. Moreover, as very small fraction of trips are classified as light-rail trips, it is assumed that other public transit modes have similar proportions in the present raw data because of low public transit usage in Dallas. Likewise, trips that have an average speed of less than 5 km h^{-1} over the course of the whole trip are classified as walking or pedestrian trips (Gong et al., 2012; Nikolić and Bierlaire, 2017) and removed from the analysis. Since the objectives of the current work lie elsewhere, extensive mode detection is not adopted.

2.4. Post-processing of segmented trips

However, there are few trips, especially by taxis or ride sharing vehicles, that have unusually long trip lengths or travel times. In reality, they are sequences of individual trips that are compounded as one long trip. This is due to the very short duration between two different trips. The occupancy of taxis tends to be high close to downtown and the idle time between the trips can be almost non-existent. On the one hand, having these types of compound trips does not influence the estimation of macroscopic variables like distance and time traveled. On the other hand, they can introduce bias in trip length distributions. Hence, care is taken to remove these types of aggregated trips for the estimation of trip lengths. This is done by comparing the network shortest path distance to the actual trip distance for a given trip. If the actual trip length is more than twice the network shortest path distance, that trip is removed from the trip length analysis. It was already shown in Yang et al. (2018a) using taxi GPS data for different sized cities that detour ratio (ratio of actual trip distance to straight line distance) follows a universal law with an asymptotic value around 1.3. It is shown in Appendix A that this ratio holds true for Dallas as well. Moreover, the ratio reduces to 1.16, when network shortest distance is used instead of straight line distance. Therefore, the present assumption of marking trips that have twice the distance of network shortest distance as unusual trips and eventually eliminating them for trip length analysis is justified. It should be noted that in the present work this technique removes a very small proportion of trips, thereby improving the trip length analysis without eliminating significant data.

If the sampling interval between consecutive records in a trip is large, depending on the speed of the vehicle, the distance covered within this interval can be significantly long. Assuming traveled distance as straight line distance between two records that are far away can introduce considerable errors. In the current work, the great-circle distance is used to take the curvature of the planet into account and it is estimated using the Haversine formula (Sinnott, 1984). It can be expressed as follows:

$$d = 2R \arcsin\left(\sqrt{\sin^2\frac{\Delta\varphi}{2} + \cos\varphi_1 \cos\varphi_2 \sin^2\frac{\Delta\lambda}{2}}\right), \tag{1}$$

where $R = 6372.8$ km is the radius of the Earth and $\Delta\varphi$, $\Delta\lambda$ are differences in latitudes and longitudes, respectively.

When two GPS coordinates are far apart, even the distance computed by the Haversine formula leads to an inaccurate estimation of traveled distance, especially when the points lie on different links in the network. One trivial solution to minimize the biases in the traveled distances is to choose the trips that have records relatively close to each other. In the first step, trips for which the consecutive records are within a radius of 500 m to each other are selected. This filtration process reduces the total number of trips to 290000, *i.e.*, 9% of the segmented raw trips. This averages to around 4700 trips per day for a relatively large area under investigation. In order to make use of all the available data, a trip enrichment method based on the map matching scheme is proposed in this work to enhance the spatial resolution of trips that have sparse records. The method employed is illustrated in the following section.

3. Trip enrichment method

The main idea behind the enrichment method is to find the network shortest path between the sparse records of a given trip. Two important tools are used in this context, namely OSMnx (Boeing, 2017), which is a Python package used to analyze road networks, and NetworkX (Hagberg et al., 2008), another Python package used to study the dynamics of road networks. The enrichment method is explained with an example in the following.

Fig. 3a shows three sample trips with sparse records from the given data. All three trips have either 3 or 4 records for relatively long trip lengths of around 10 km. The result is a very poor spatial resolution of the trajectory of the trip, which is shown in Fig. 3a. It is also evident that the trips either start or end close to the downtown region of Dallas and end in the suburbs of the city, which suggest they are realistic trips across the city that span a considerable amount of time. Using Haversine distance between the successive data points of each trip introduces a huge approximation in the traveled distances. Furthermore, the traveled time estimation is unaffected due to the sparse records. The combination of these two phenomena can influence the shape of the MFD and result in poor trip length estimation. Hence, it is desirable to map the trajectory of the trip with respect to the underlying road network as closely as possible.

In the current work, trips are enriched using the spatial geometry of the network. OSMnx contains the information of the whole network in the form of links and nodes. For each trip, the distance between successive records is estimated. If the distance is longer than a threshold, defined *a priori*, the locations of the nearest nodes close to the GPS positions of those records are obtained. The threshold distance can be defined depending on the average block size of the road network. In the current case of Dallas, the block size in the downtown area is smaller than that in the suburban areas. It is possible to define different threshold distances based on the location of the trip in the current framework. For the sake of simplicity, a constant threshold distance of 200 m is assumed in the present work. Once the locations of the nodes are obtained between the sparse data records, the network shortest path between these

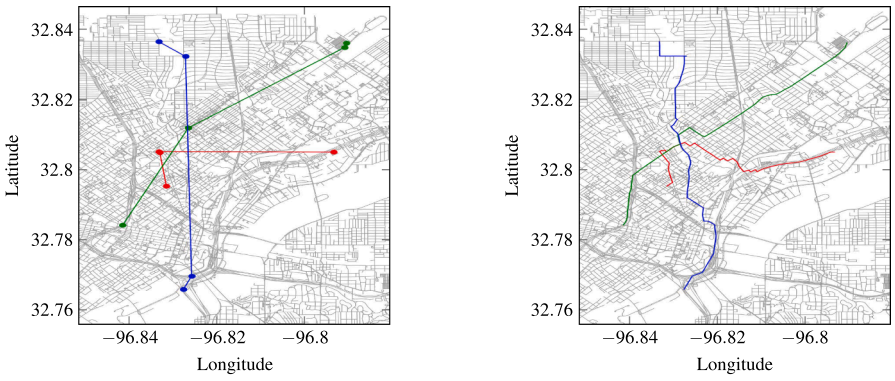

Fig. 3. Trip enhancement method. (a) Original traces of low resolution trips. (b) Traces of the trips after the enrichment method.

two nodes is computed using NetworkX library. The network shortest path is represented as the sequence of nodes at each intersection between the given two nodes. Consequently, the GPS locations of the nodes in the network shortest path are added to the trip between these two sparse records. Fig. 3b presents the traces of the trips considered after computing the network shortest paths between the sparse records. It can be clearly seen that the trajectory of each trip is matched to the network after the enrichment process.

Another trivial method used to enrich the trips is to estimate the network shortest path between the origin and destination locations of each trip. However, this method fails to capture the longer paths that users tend to take during peak hour congestion periods. On the contrary, the trip enrichment scheme proposed in the current work keeps the original structure of the actual trip, while only adding the network shortest path between the records that are sparsely placed. Hence, this can be considered as the closest approximation to the actual path that the user had taken. The main advantage of this method is that the enriched trips conform to the actual network, which can be observed in Fig. 3b. The main limitation is that the travel time on the computed network shortest path might not be consistent with the actual travel time between these two records. Using the speed data on individual links, it is possible to estimate the travel time for the computed network shortest path. If this travel time is inconsistent with the actual travel time, *i.e.*, the difference between the timestamps of the records, another network path that satisfies the actual travel time must be estimated. However, this demands considerable computational resources due to the huge mass of data in the present case. Moreover, it is shown in the subsequent sections that the errors of the present framework are within an acceptable tolerance.

Whenever the trajectory of a trip is enhanced between two sparse records, timestamps are also interpolated to match the spatial data. The timestamps are interpolated based on the average speed between the two records in the original trajectory. Hence, this method transforms the sparse spatial and temporal data into dense data, thereby improving the overall accuracy of traveled distances. Fig. 4a and b present a sample set of traces of trips before and after the enrichment method, respectively. It is clear from the plots that the original trips have sparse records and that enhancing the trips results in high resolution traces. It is noteworthy that the trips in Fig. 4a and b are randomly sampled and do not correspond to the same set of trips.

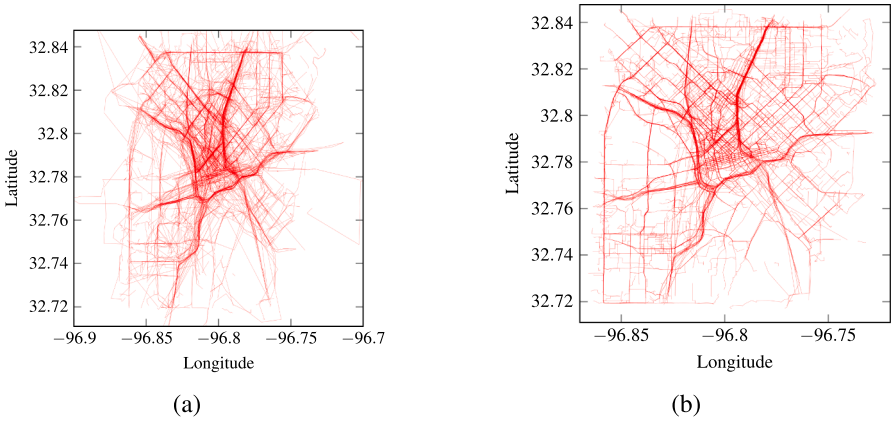

Fig. 4. Trip enhancement method. (a) Randomly sampled original traces. (b) Traces after the enrichment method.

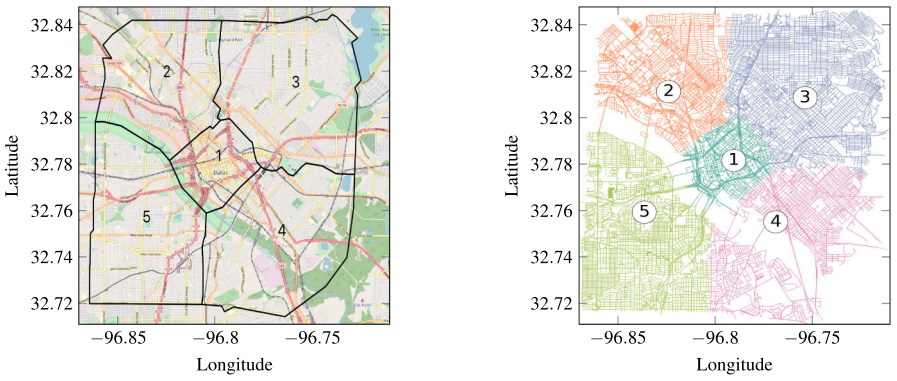

Fig. 5. Dallas network partition. (a) Dallas network partition ©OpenStreetMap 2019. (b) Link level representation of reservoirs.

4. Computation of macroscopic variables

4.1. Partitioning of the network

The partitioning of the network considered into homogeneous reservoirs is the first step in the estimation of macroscopic traffic variables. A prerequisite of obtaining a well-defined MFD is to partition the network into homogeneous subnetworks (Geroliminis and Sun, 2011). Partitioning algorithms have been proposed in the literature, based on the properties of links (Ji and Geroliminis, 2012; Saeedmanesh and Geroliminis, 2016) and on traffic data (Ambühl et al., in press). As the primary objective of the current work is to propose a methodology to estimate the parameters of MFD-based models using mobile phone data, a rather simple partitioning method is used. However, it should be noted that the present framework can be used with any of the partitioning schemes proposed in the literature. Moreover, in order to observe and study congestion patterns during peak periods, it is imperative to consider a more sophisticated partition. The network considered is divided into 5 reservoirs, as shown in Fig. 5. The rationale behind the partition is to have one reservoir for the downtown region and divide the suburbs around downtown into similar sized reservoirs. They are divided in such a way that all of them have an equal number of free-ways. It is important to place the boundaries of the reservoirs in-between road networks and along roads that should be avoided. Generally, the boundaries of partitions are placed along the roads, where adjacent zones share their widths. Since GPS data have uncertainties, placing a boundary on the road might result in trips that alternate between two adjacent reservoirs, even though they belong wholly to one of these neighboring reservoirs. This can introduce biases in the estimation of MFDs and trip length distributions for these two adjacent reservoirs. This effect is more pronounced if boundaries are placed along major arterials and freeways, where higher mean flows are observed. These types of alternating trips can be avoided by placing boundaries within the blocks of the road network and it is the reason for the irregularity of the boundary in the present partitioning.

4.2. Error estimation of trip enrichment scheme

In this section, the error in traveled distances introduced by the proposed trip enrichment method is estimated. In order to do so, trips from the raw data with high spatial resolution, *i.e.*, trips that have successive records within a radius of 200 m from their neighbors, are chosen for each OD pair. Since the records of these trips are relatively close to each other, they conform with the network topology. This set of trips for each OD pair is considered as the reference set, say $\{L_f\}$, to estimate the error in the traveled distances. The idea is to transform these high-resolution reference trips to low resolution ones, say $\{L_c\}$, by randomly removing the intermediate records of each trip. In this work, 70% of records are removed for each reference trip while keeping the origin and the destination records unchanged. The mean sampling interval of the high resolution reference trips is 9.5 s, whereas it increases to 42 s, a fivefold increase, after the removal of data records. Moreover, the sampling distance, *i.e.*, distance traveled within a sampling interval, is a better indicator for studying the accuracy of the trip enrichment method. Fig. 6 presents the probability distribution of sampling distances of the original high resolution trips and low resolution trips after removing the records. It is evident that data records are further apart in the low-resolution data. The mean sampling distance in the high and low resolution trips are 93 m and 313 m, respectively.

Now, these low-resolution reference trips are enriched using the enrichment method discussed in the previous section to obtain the map matched trip trajectories. The enriched trips are denoted by $\{L_e\}$. As the variable of interest in the current work is traveled distance, the error is estimated based on the difference in the trip distances between high resolution reference trips and enriched trips. Table 1 presents the percentage of relative Root Mean Square Error (RMSE) of all the reference trips for the major OD pairs. It is clear from the error values that the enrichment scheme is very accurate in terms of traveled distances. Only the OD pairs for which there are more than 1000 reference trips are presented in the table; the error values for the remaining OD pairs are also between range 5 to 7 %. Table 2 shows the average trip lengths of the reference ($\overline{L}_f$), low resolution ($\overline{L}_c$) and enriched ($\overline{L}_e$) trips inside each

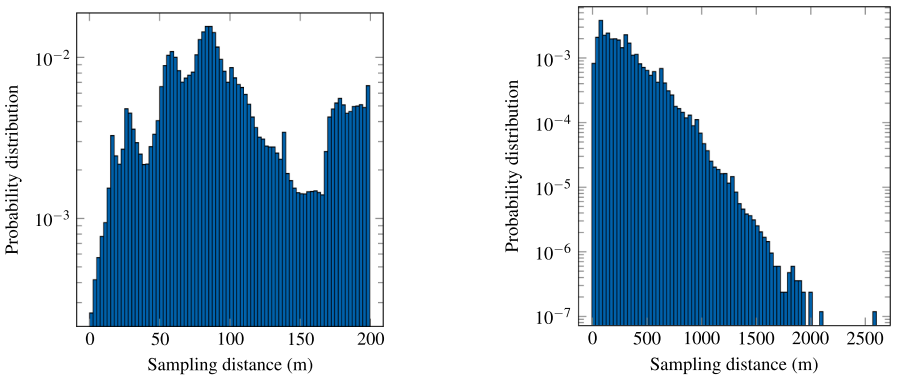

Fig. 6. Sampling distances for high and low-resolution trips used in error estimation of trip enrichment technique. (a) Original high resolution trips. (b) Low resolution trips after removal of records.

Table 1
Relative RMSE in % of the trip enrichment method for the reference trips.

OD	Error	# trips	OD	Error	# trips
1–1	7.6	3629	3–1	6.2	1802
1–2	5.6	1052	3–3	4.9	10182
1–3	5.5	1766	4–4	5.5	1429
2–1	5.8	1321	5–5	5.0	4742
2–2	5.7	9403			

Table 2
Average trip lengths of the reference trips inside each reservoir.

Res.	$\overline{L_f}$ (m)	$\overline{L_c}$ (m)	$\overline{L_e}$ (m)	$\left(\frac{\lvert L_f - L_e \rvert}{L_f}\right)$ in %
1	1097	858	1068	2.6
2	1414	1254	1385	2.0
3	1530	1323	1515	0.9
4	1480	1302	1448	2.1
5	1589	1380	1545	2.7

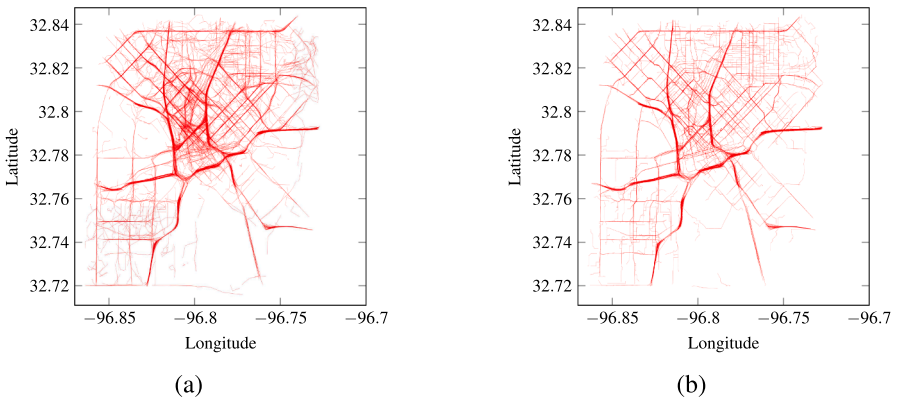

Fig. 7. Randomly sampled reference low resolution and enriched trips for all OD pairs. (a) Reference low resolution trip samples for all OD pairs. (b) Reference enriched trip samples for all OD pairs.

reservoir. It is evident that low resolution trips under-predict trip lengths by range 10 to 20 % on average in all the reservoirs. In addition, the enrichment method yields very good approximations of actual trip distances, where errors less than 3 % are observed. Thus, it can be concluded that the proposed trip enrichment technique gives very good approximations of actual traveled distances. Fig. 7 shows the sample reference low resolution and enriched traces. Although the major arterials are well represented in the low resolution traces, there is considerable scatter between the arterials. This is corrected in the enriched traces, where the trip trajectories conform well with the underlying network. Hence, it can be concluded that the errors introduced from the proposed trip enrichment scheme are within the acceptable limits for the present purpose of estimating macroscopic variables.

4.3. Estimation of macroscopic variables and penetration rates

According to the generalized definitions of Edie (1963) and Nagle and Gayah (2014), average network density (k) and flow (q) can be expressed as,

$$k = \frac{\sum_{i=1}^{P} TT_i/\rho}{L_n\,\Delta T} \text{ and } q = \frac{\sum_{i=1}^{P} TD_i/\rho}{L_n\,\Delta T}, \tag{2}$$

where TT_i and TD_i are the time traveled and distance covered on a link l, respectively, L_n is the total network length, ΔT is the aggregation interval, N is the total number of links in the network and ρ is the penetration rate and defined as the number of probe vehicles to the total number of vehicles in the network. In the present work, an aggregation interval of 15 min is used. It is clear from eq. (2) that the density is computed using Total Traveled Time (TTT) and the flow is estimated by Total Traveled Distance (TTD) in the network. In eq. (2), the sum is calculated over the total number of probe vehicle trajectories P.

It is necessary to estimate the penetration rate (ρ) of the vehicles to compute network-wide macroscopic variables. A constant penetration rate as in eq. (2) can be estimated by using the traffic counts from fixed loop detectors and trips that pass at the corresponding locations in the data. However, this method gives a mean penetration rate that does not account for network heterogeneity and thus, it might not be accurate enough. It was already concluded that using an OD specific penetration rate is important for estimating an accurate MFD (Du et al., 2016). Generally, higher penetration rates are observed in the downtown area compared to suburban areas. In the same way, higher rates are normally observed at peak hours compared to off-peak hours. Thus, it is important to consider the spatial and temporal variation of penetration rates.

The current work proposes two different types of penetration rate, namely OD specific (ρ_{od}) and origin specific penetration rates (ρ_o). It is possible to estimate the trip OD matrix for a given aggregation interval from mobile phone data based on the departure time of each trip. The trip OD matrix is estimated for each aggregation interval and each day separately. Let $N^I_{od,p}$ be the number of probe trips from data between origin o and destination d, starting within the aggregation interval I. Similarly, $N^I_{od,n}$ is the total number of trips in the network within that interval I. Now, the OD specific penetration rate at aggregation interval I can be defined as,

$$\rho^I_{od} = \frac{N^I_{od,p}}{N^I_{od,n}}. \tag{3}$$

Similarly, let $N^I_{o,p}$ and $N^I_{o,n}$ be the total number of trips originating from origin o to all destinations for a given interval I in the data and the actual network, respectively. Then, the origin specific penetration rate can be expressed as,

$$\rho^I_o = \frac{N^I_{o,p}}{N^I_{o,n}} \equiv \frac{\sum_{d=1}^{r} N^I_{od,p}}{\sum_{d=1}^{r} N^I_{od,n}}, \tag{4}$$

where r is the total number of macroscopic reservoirs in the network. Estimating probe trips between OD pairs for each aggregation interval from data is straightforward as trip trajectories and corresponding timestamps are available. Thus, to compute the proposed time-dependent OD specific and origin-specific penetration rates, it is necessary to have information on the dynamic OD matrix of the entire network.

4.4. Estimation of dynamic OD matrix

In the current work, the dynamic OD matrix is estimated by fusing data from two different sources. The available data sources in this context are:

- A static OD matrix provided by the North Central Texas Council of Governments (NCTCOG, 2019) for the morning peak period, *i.e.*, 06:30 AM to 08:59 AM. This matrix contains the cumulative number of vehicles traveling from one region to another. These regions are defined by the NCTCOG and they are typically smaller than the partition presented in Fig. 5.
- The second source of data is obtained from loop detectors. The data of total traffic counts of loop detectors placed all over the Dallas city network are made available by the NCTCOG. A few loop detectors that cover the area considered are chosen and the traffic counts of these loop detectors are used in the analysis.

Since daily traffic (total counts for a 24-h period) on each loop detector can vary widely, traffic counts on each loop detector are normalized by the daily traffic of that particular loop-detector. For instance, if loop detector LD_i has a traffic count tc^I_i in an

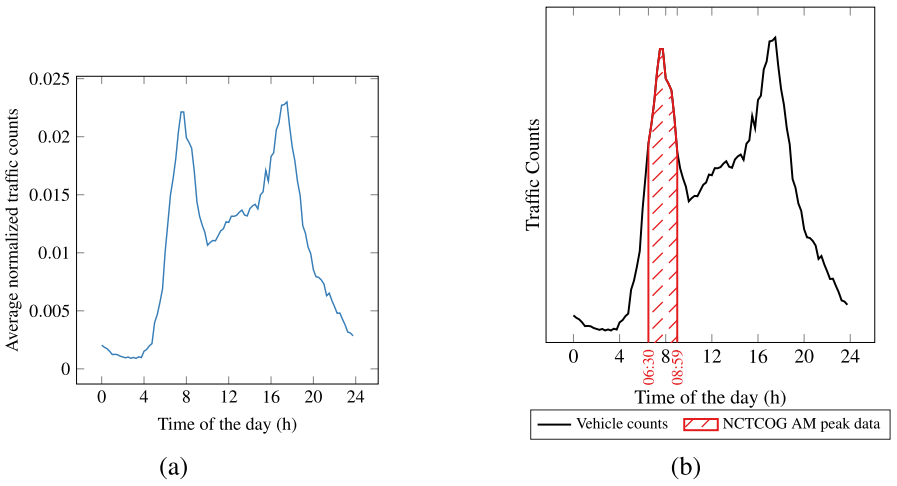

Fig. 8. Dynamic OD matrix calibration method. (a) Mean of normalized traffic counts from LDD. (b) Expanding the static OD matrix into a dynamic one.

aggregation interval I, then the normalized traffic count will be $\frac{\mathrm{tc}_i^I}{\sum_I \mathrm{tc}_i}$, where daily traffic ($\mathrm{dt}_i$) is defined as $\sum_I \mathrm{tc}_i$. Using these normalized traffic counts on each loop-detector, an average normalized count in each aggregation interval is estimated, *i.e.*, $\bar{\mathrm{tc}}^I = \sum_i \frac{\mathrm{tc}_i^I}{\sum_I \mathrm{tc}_i}$. The average normalized traffic counts curve is used as a proxy for the daily demand pattern in Dallas.

Fig. 8b shows the computed average normalized traffic counts curve for a typical day. The morning peak is observed between 08:00 and 08:30 and the evening peak is observed between 17:00 and 17:30. The objective is now to transform the available static OD matrix into a dynamic one using this normalized traffic count curve. Let us consider that for a given OD pair *od*, dynamic traffic counts follow the curve presented in 8b. Notice that the traffic counts here correspond to total vehicle counts within an aggregation interval. The NCTCOG OD matrix data correspond to the shaded area in the graph. In other words, the area under the shaded part gives the cumulative number of vehicles moving between the considered OD pair. Let us consider that TC_{od}^I is the traffic count at an aggregation interval I. Since it is already assumed that the normalized traffic count of all OD pairs follows the curve presented in 8b, it is imperative that $\frac{\mathrm{TC}_{od}^I}{\sum_I \mathrm{TC}_{od}} = \bar{\mathrm{tc}}^I$. Since $\bar{\mathrm{tc}}^I$ is already known and TC_{od}^I is known for aggregation intervals between 06:30 and 08:59, it is possible to compute the daily traffic of the given OD pair, *i.e.*, $\mathrm{DT}_{od} = \sum_I \mathrm{TC}_{od}$. It can be expressed as,

$$\mathrm{DT}_{od} \equiv \sum_I \mathrm{TC}_{od} = \frac{\sum_{I\in\{06:30,08:59\}} \mathrm{TC}_{od}^I}{\sum_{I\in\{06:30,08:59\}} \bar{\mathrm{tc}}^I}. \tag{5}$$

Once the daily traffic of the OD pair is estimated, the traffic counts at any aggregation interval can be computed as $\sum_I \mathrm{TC}_{od}^I \bar{\mathrm{tc}}^I$. This method fuses the static OD matrix data and the dynamic traffic count curve to estimate the dynamic OD matrix for a typical day. Table 3 gives the total number of vehicles for the macroscopic OD pairs, *i.e.*, vehicle counts from one reservoir to another for a 24 h period.

As the OD matrix typically provides trip production and trip attraction data, penetration rates should be computed based on the departure times of trips. Consider a trip i starting within an aggregation interval I between the OD pair *od* with a traveled time and a distance of TT_i and TD_i, respectively. The expanded travel distance (TD_i^e) and the expanded travel time (TT_i^e) for the whole network using two types of penetration rate for that trajectory can be expressed as,

$$\begin{aligned} TT_i^{e,od} &= \frac{TT_i}{\frac{I_p}{\rho_{od}}} \quad \text{and} \quad TD_i^{e,od} = \frac{TD_i}{\frac{I_p}{\rho_{od}}}, \\ TT_i^{e,o} &= \frac{TT_i}{\frac{I_p}{\rho_o}} \quad \text{and} \quad TD_i^{e,o} = \frac{TD_i}{\frac{I_p}{\rho_o}}, \end{aligned} \tag{6}$$

where suffixes/prefixes *od* and *o* represent expansion done by OD and origin specific penetration factors, respectively, and I_p is the

Table 3
Total traffic counts ($\times 10^4$) for 24 h period.

Reservoir	1	2	3	4	5
1	7.9	3.4	2.6	0.5	0.7
2	4.8	7.2	1.8	0.2	0.4
3	7.0	4.4	6.5	0.4	0.2
4	1.7	0.9	0.7	0.7	0.1
5	4.0	3.0	0.7	0.2	2.2

Table 4
Total network length in km for each reservoir.

Reservoir	Length of the network (km)
1	324.8
2	414.8
3	431.2
4	309.6
5	522.3

aggregation interval corresponding to the *departure time* of trip *i*. Following eq. (2), k and q can be estimated. Finally, the total length of the network L_n is estimated per reservoir to compute mean density and mean flow. Only major roads are considered in the estimation of total network length. The flow on residential roads is typically very low, hence they are neglected. Moreover, total network length is used to normalize only the Total Traveled Time and Total Traveled Distance to express them in terms of mean flow and density. Table 4 shows the resulting lengths of the road networks per reservoir.

As it will be shown in Section 5, both OD specific and origin specific penetration rates yield very similar MFD shapes. Thus, if the OD matrix data is not available, origin specific penetration rates can be estimated using the census data. The population data is readily available in most cases and can be used as a proxy for the trip production for each aggregate zone. This population data, along with the LDD, can be used to build a dynamic trip production matrix. Thus, census data can be used in place of OD matrix data in the proposed framework to estimate the time dependent penetration rates.

5. Estimation of penetration rates and MFDs

5.1. MFD estimation

The mean densities and flows for each aggregation interval for each day are estimated using eqs. (2). Only data from weekdays, which are 43 in total, are considered in the analysis to estimate a stable and reproducible MFD. A mean MFD for each reservoir can be estimated by considering *all* weekdays. However, specific events like accidents, road works, *etc.* or weather conditions can influence the shape of MFD on certain days. Since the information of historical events is not available, the data is filtered based on the estimated MFD shape. Another important factor to consider in this context is the phenomenon of hysteresis in the MFD for the urban networks (Buisson and Ladier, 2009; Leclercq and Paipuri, 2019). It is normal to observe hysteresis loops in the MFD due to network heterogeneity, demand pattern and driver behavior, *etc*. However, the loading of the network from near empty state, which is usually observed during late night hours, is more stable and reproducible. Hence, a parabolic curve is fitted for the MFDs estimated for each weekday for the data point ranging from midnight until the morning peak hour, i.e. 08:00 AM in the present case. Only days that show a similar MFD shape are chosen to compute the mean MFD for each reservoir. In the present case, it is observed that on average 18 days present similar MFD shapes. However, it should be noted that these 18 days can be different days for different reservoirs. Day-to-day MFDs are shown for reservoirs 1, 2 and 3 in Appendix B for selected days.

Fig. 9 shows the estimated flow MFDs for the reservoirs, where the macroscopic variables are computed by the OD specific penetration rate. Firstly, it can be noticed that all the MFD curves are relatively stable with very little scatter. Reservoir 1 undergoes the highest flow rate among all the reservoirs in the region considered. From Fig. 5a, it is clear that this reservoir corresponds to the downtown Dallas area and hence, a higher mean flow is observed. Another important inference to be made from the MFD plots is the presence of clockwise hysteresis loops in the MFDs. Different colors of the data points correspond to the different times of day to differentiate the loading and unloading phases of the morning and evening peak hours. It can be noticed that data points from 12:01 AM to 08:00 AM in estimated MFDs have higher flow values than others. This is the result of the hysteresis phenomenon. It can be seen in the speed MFDs in Fig. 10 that mean speed tends to be higher during the network loading from midnight to the morning peak hour than the rest of the day. That trend is translated on the production MFD plane, where higher flows are observed during the morning loading period.

Reservoirs 1 and 3 in Fig. 9a and c, respectively, exhibit clearer hysteresis during the morning peak hour. Using Google typical traffic data, it is noticed that there are several internal bottlenecks on major arterials and freeway networks in these two reservoirs 1 and 3, leading to severe local congestion. Previous studies (Leclercq and Paipuri, 2019) showed that multiple active bottlenecks in the network saturation state trigger the hysteresis patterns in the network. Thus, the hysteresis loops observed in reservoirs 1 and 3 might be the result of a similar mechanism. It is also observed that the freeway network in reservoir 2 is less congested and hence, a smaller hysteresis loop is noticed. Finally, reservoirs 4 and 5 shown in Fig. 9d and e experience lower mean flows and densities and no hysteresis loops. This may be due to less congestion in those regions. Other reasons include the lack of robust data in these zones and the presence of large urban spaces for leisure activities.

Fig. 10 presents the mean speed MFDs using the OD specific penetration rate. The hysteresis phenomenon can be clearly observed in the speed MFDs, while in Fig. 10a and c, the mean speed during loading is clearly higher than during unloading in the morning peak hour. The estimated speed MFD in reservoir 4, which is shown in Fig. 10d, has comparatively large scatter. This is due to sparse phone data in this reservoir. It can also be noted that the mean speed in reservoir 4 does not vary as much as other reservoirs. This is due to the linear relationship that is observed between mean flow and mean density in Fig. 9d. It is logical that free flow speeds in reservoirs 2 to 5 should be higher than in reservoir 1, which is a downtown area. It can be observed from the plots that the free-flow

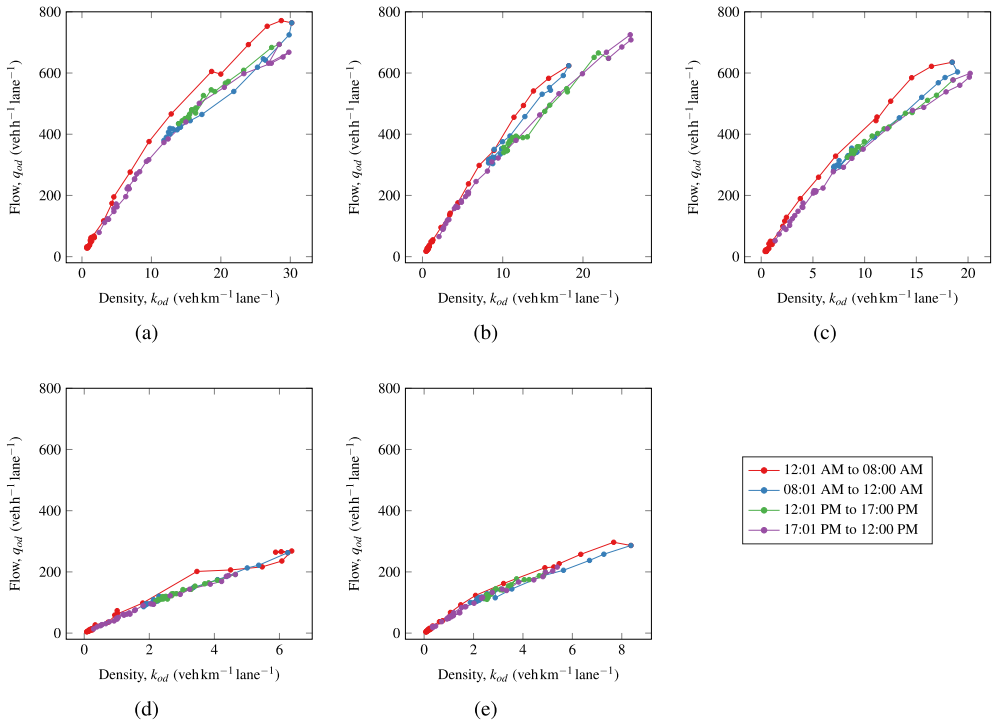

Fig. 9. Flow MFD estimates using OD specific penetration rate. (a) Reservoir 1. (b) Reservoir 2. (c) Reservoir 3. (d) Reservoir 4. (e) Reservoir 5.

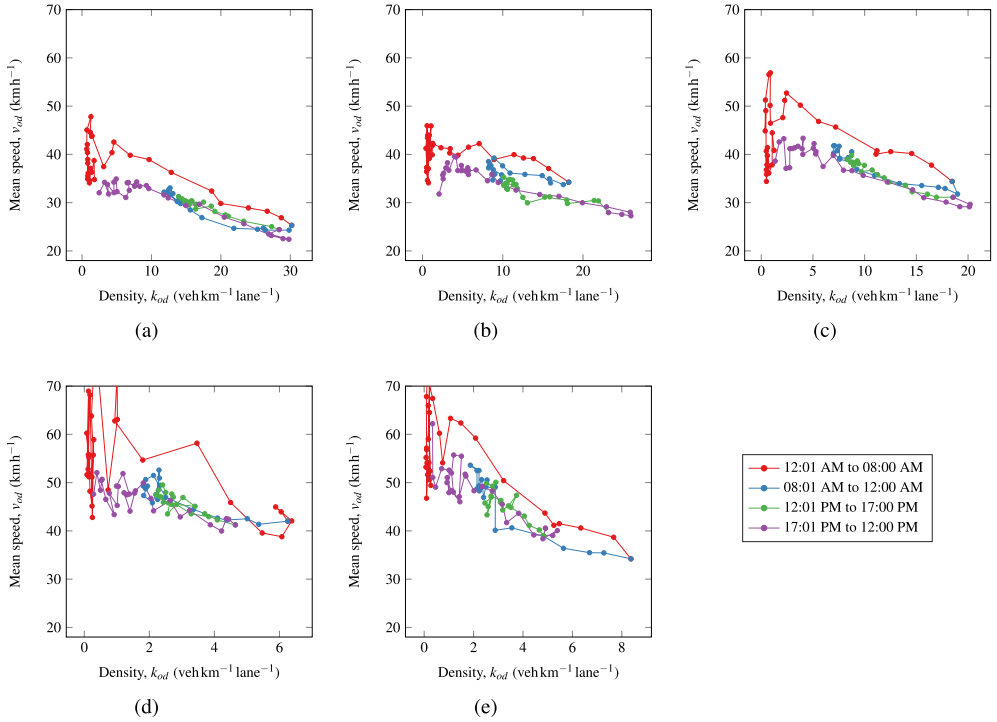

Fig. 10. Speed MFD estimates using OD specific penetration rate. (a) Reservoir 1. (b) Reservoir 2. (c) Reservoir 3. (d) Reservoir 4. (e) Reservoir 5.

speed of reservoirs 2 to 5 is higher than that of reservoir 1 and hence the estimated MFDs are verified qualitatively.

Fig. 11 presents the estimated flow MFDs using origin specific penetration rate for computing macroscopic variables. Firstly, it is clear that the MFDs are qualitatively and quantitatively very similar to those presented in Fig. 9, where the OD specific penetration rate is used in the computation of macroscopic variables. The hysteresis loops observed in reservoirs 1 and 3 in the previous case are

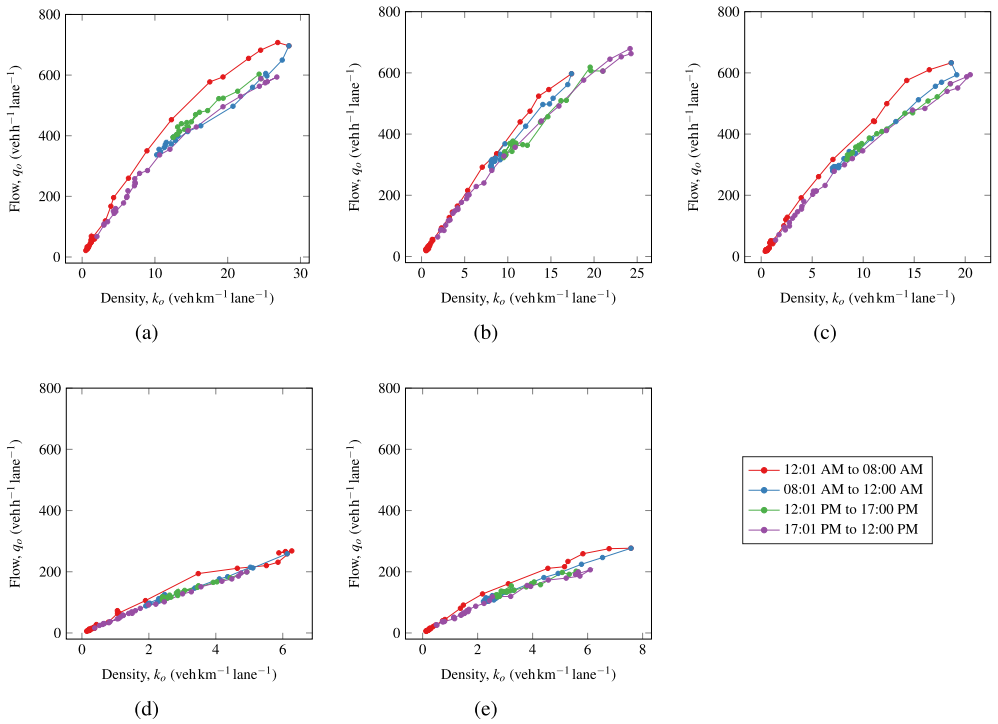

Fig. 11. Flow MFD estimates using origin specific penetration rate. (a) Reservoir 1. (b) Reservoir 2. (c) Reservoir 3. (d) Reservoir 4. (f) Reservoir 5.

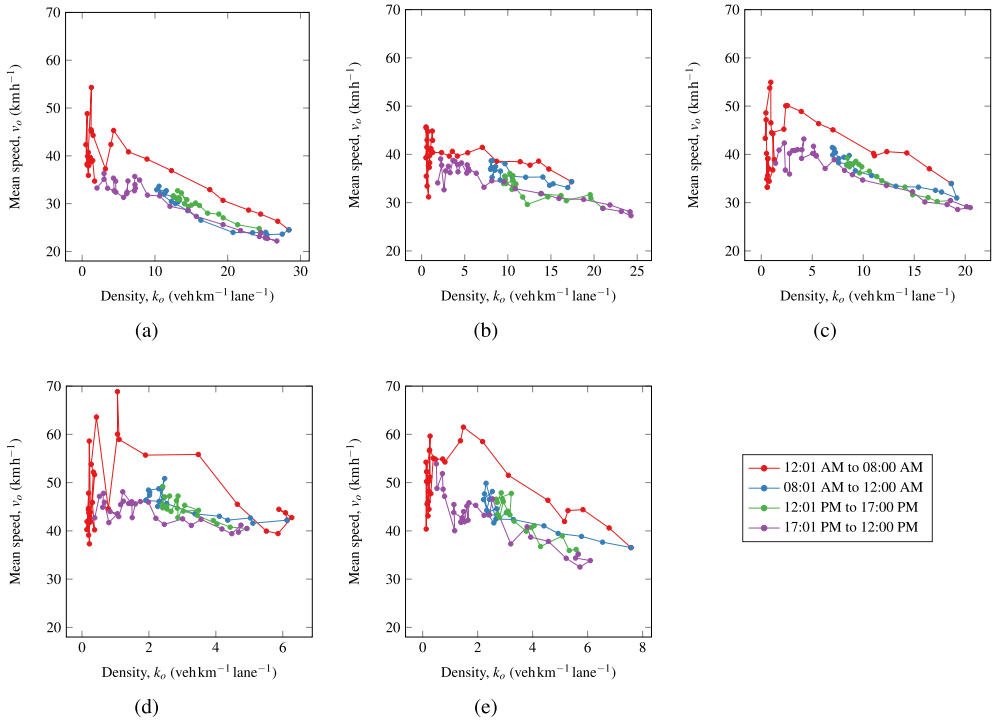

Fig. 12. Speed MFD estimates using origin specific penetration rate. (a) Reservoir 1. (b) Reservoir 2. (c) Reservoir 3. (d) Reservoir 4. (f) Reservoir 5.

also observed in the present case. However, scatter in the reservoir 5 in Fig. 11e is comparatively larger than the scatter in Fig. 9e. This is true not only for reservoir 5 but also for all the reservoirs. This is clearly observed in the speed MFD plots shown in Fig. 12. Although all the plots show a good relationship between mean speed and density, it is evident that the scatter in the present case is relatively larger than in the previous case. This is expected as the OD specific penetration rate is more accurate than the origin

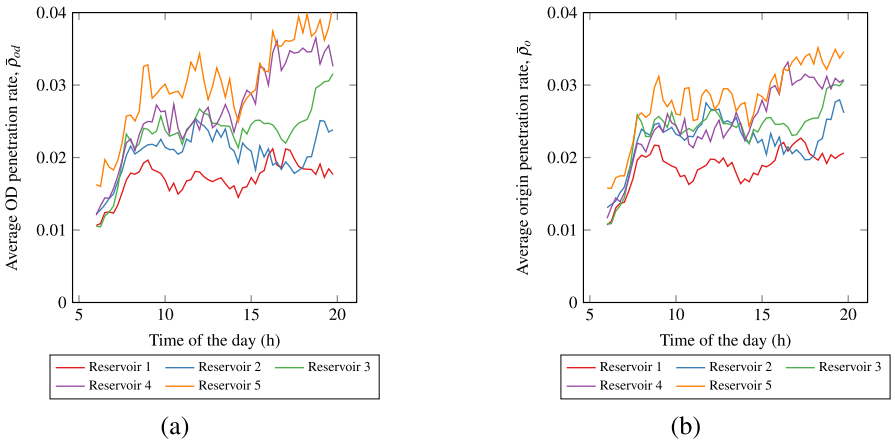

Fig. 13. Evolution of average penetration rates for each reservoir. (a) OD specific. (b) Origin specific.

specific counterpart. Expanding the macroscopic variables without considering the destination might result in overestimation in a few OD pairs and underestimation in the rest. This can contribute to the scatter in the MFD, observed in Figs. 11 and 12. Despite the presence of scatter, both the flow and speed MFDs show a reasonable correlation between the macroscopic variables.

5.2. Discussion on penetration rates

Fig. 13 shows the average penetration rates of probe vehicles from LBS data for each reservoir. It should be noted that according to the present framework, penetration rates within the same aggregation interval are not the same and they depend on the departure time of the trip as presented in eq. (6). Hence, multiple penetration rates are possible for the same aggregation interval as multiple trips can co-exist during that period. Average penetration rates are computed by comparing the expanded macroscopic variables estimated from LBS data to the OD matrix presented in Table 3. The estimated penetration rates for different days are again averaged to obtain an average trend of variation of penetration rates during a typical day scenario. In order to obtain a representative trend, only days that show similar MFD shape are considered to estimate average penetration rates. This is done for both OD specific and origin specific penetration rates. The first inference from the plots is that both the proposed penetration rates show a very similar trend, as expected. It can be observed that the peak penetration rates are obtained during morning and evening peak hours. This phenomenon is can be seen clearly in reservoir 1, where two peaks, one in the morning and another in the evening, can be observed. It is also evident that the variation of the penetration rate within a day cannot be neglected, and that using a mean penetration for the whole day can lead to erroneous results.

In order to show the importance of considering time dependent penetration rates, MFD is estimated using a constant penetration rate for a whole 24 h period. Fig. 14 presents the estimated MFDs using time averaged constant and time dependent penetration rates for reservoir 1. It is clear from the plots that the MFD estimated using constant penetration has less scatter than its counterpart. In other words, a smaller hysteresis loop is observed in Fig. 14a. At the same time, significant hysteresis pattern is observed in Fig. 14b, where MFD is estimated using time dependent penetration rate. It can be inferred that using a constant penetration rate can de-emphasize the presence of hysteresis with this particular type of data. The is due to the differences in penetration rates during

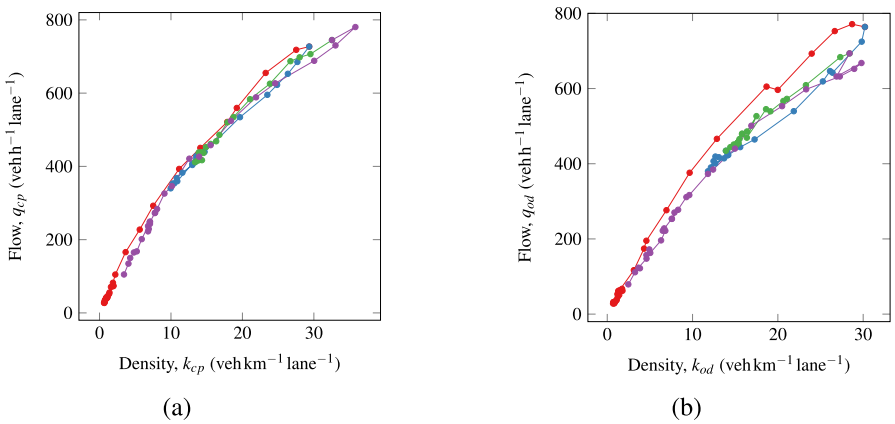

Fig. 14. Comparison of MFDs using constant and time varying penetration rates for reservoir 1. (a) Time-averaged constant penetration rate. (b) Time-dependent OD specific penetration rate.

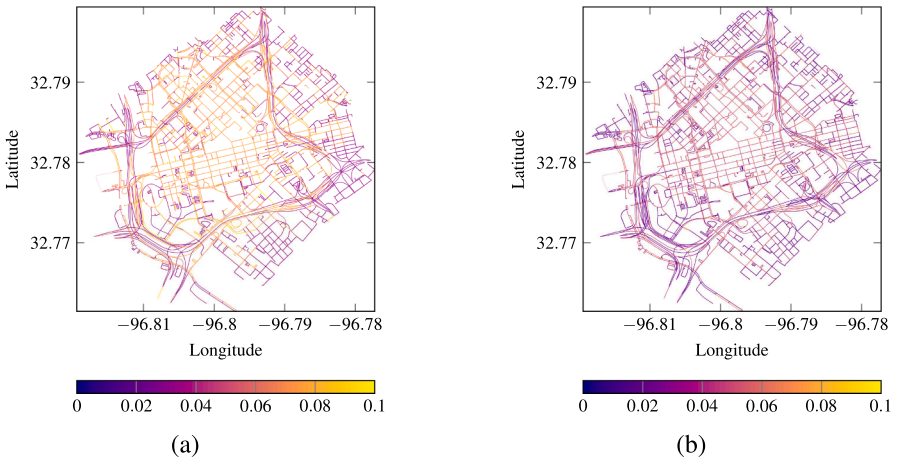

Fig. 15. Spatial distribution of OD specific penetration rates during peak and off-peak hours. (a) Morning peak hour: 08:45 AM – 09:00 AM. (b) Off peak hour: 13:45 PM – 14:00 PM.

network loading and unloading phases. Thus, care should be taken while drawing inferences from MFDs estimated by such constant penetration rate methods. Therefore, in order to accurately estimate the MFD shape using mobile phone data, it is necessary to take into account the time dependency of penetration rates, especially during network loading and unloading periods.

Besides the temporal variation, the spatial variation of penetration rates is also an important factor to be considered in estimating accurate MFDs. Fig. 15 presents the variation of time-averaged OD specific penetration rates in the downtown area of Dallas. Two different aggregation intervals, one at peak hour and one at off-peak, are shown in Figs. 15a and b, respectively. It is evident that the penetration rates are higher during the peak hour period than during the off-peak period. It can be seen in Fig. 15a that few segments of the ring road have lower penetration rates. This is due to the fact that various freeways run very close to each other and that the uncertainty on GPS coordinates can result in several freeways having a very low number of GPS records while their neighbors have high numbers of records. As macroscopic variables are estimated at the network-wide level, this bias on certain links has no influence on the calibrated parameters. It is clear from Fig. 15b that the links that underwent high penetration rates in the peak hour period have moderate to low levels of penetration during the off-peak period. It is imperative to take into account that assuming a constant penetration rate per reservoir can introduce significant errors in the estimated parameters.

Finally, discussion on penetration rates is concluded by presenting a study on the level of aggregation needed to estimate representative penetration rates. The results presented until now assume the network partition presented in Fig. 5 to estimate penetration rates and MFDs. However, the static OD matrix obtained from NCTCOG assumes spatial aggregation at a finer level, with a total number of 467 clusters. Thus, information on OD flows at a finer level of aggregation is available. This finer OD flow information can be used to estimate penetration rates. Since penetration rates are associated with a given trip and not underlying partition, MFDs can be computed for any network partition assumed. In other words, two different partitions can be defined: (i) an initial partition to compute penetration rates and (ii) a secondary partition to estimate the MFDs in each sub-network using the penetration rates obtained from the first partition. These two partitions can be independent and there are no strong geometric constraints necessary between them. The objective in the present context is to study how the spatial aggregation in so-called initial partition can influence the shape of the estimated MFD in the secondary partition. It should be noted that MFDs are always estimated for the partition presented in Fig. 5.

Three different levels of spatial aggregation are considered here, which are shown in Fig. 16. The spatial clusters provided by the NCTCOG are merged into 50, 100 and 300 using the k-means clustering algorithm. The trip OD matrix and the actual OD matrix are

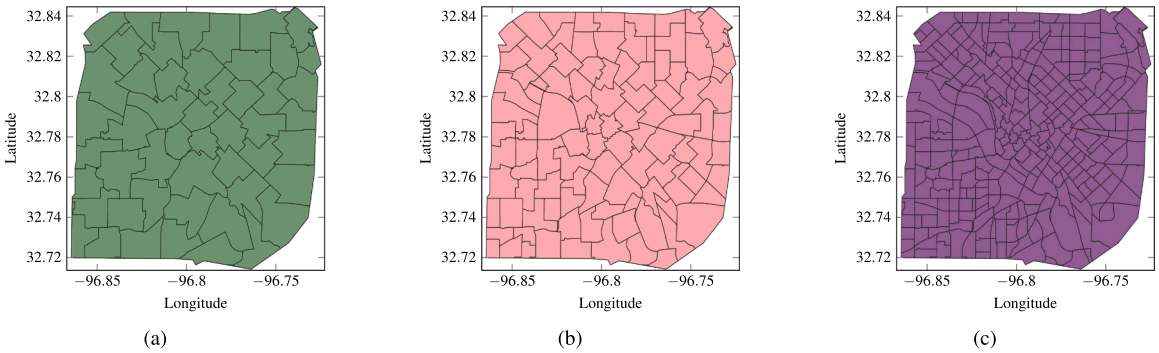

Fig. 16. Different spatial aggregations considered. (a) 50 clusters. (b) 100 clusters. (c) 300 clusters.

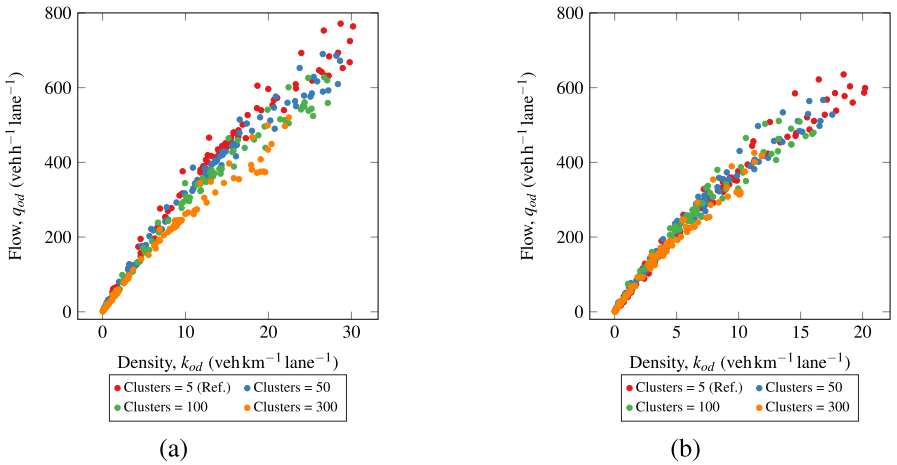

Fig. 17. Flow MFD estimates using the OD specific penetration rate. (a) Reservoir 1. (b) Reservoir 3.

estimated for each level of aggregation and, finally, the penetration rates are computed. The same approach presented in Section 4.3 is used to compute penetration rates and macroscopic variables. The only difference is that two different partitions are used in the present case. Fig. 17 presents estimated MFDs for reservoirs 1 and 3 using OD specific penetration rates. The reference partition in Fig. 5 is indicated, also plotted in the graphs as 5 clusters, to compare results. It is clear from the plots that each spatial cluster gives different MFDs. Only reservoirs 1 and 3 are chosen to present results as they exhibit clear hysteresis loops and thus the differences in the results are more obvious. An important inference here is that as the number of clusters increases, the peak density and flow values decrease. The difference in peak flows between reference the 5 cluster cases and 300 cluster cases is approximately 300veh $h^{-1}lane^{-1}$ for reservoir 1. A similar trend is noticed for reservoir 3 as well, where the smallest MFD is obtained with the finest spatial clustering. This is due to an unaccounted flow that is lost because of finer spatial clustering. This can be better explained by studying the properties of actual and trip OD matrices.

From Eq. 3, it is clear that if OD trips at a given aggregation interval, $N^I_{od,p}$, is zero, the penetration rate is zero provided that the actual OD trips, $N^I_{od,n}$, is non-zero. This signifies that there are trips between a given OD pair in reality, but these OD trips are not captured from the data and consequently, the flow between that OD pair at that aggregation interval is not captured. In other words, if there are no trips segmented from the data between a given OD pair, the flow is implicitly assumed as zero between that OD pair. This is the reason why lower flows are obtained for finer spatial clusters in Fig. 17. More and more OD pairs in the trip OD matrix have zero trips as the size of the clusters decreases. Table 5 presents densities of actual and trip OD matrices for different levels of aggregation. The density of a matrix is defined as the ratio of the number of non-zero elements to the total number of elements. Thus, density gives information on non-zero trip OD pairs. The densities of the trip OD matrix are averaged for all the selected days (between 06:00 AM to 20:00 PM) to estimate the mean density. In the case of 300 spatial clusters, the density of the trip OD matrix is very low while the density of the actual OD matrix is close to one. This implies that there are non-zero flows between all the OD pairs in reality, but the trip OD matrix has only one non-zero OD pair for every 100 OD pairs. As the number of clusters decreases, the density of the trip OD matrix increases and for the reference case of 5 clusters, the density of the trip OD matrix is unity. Thus, the reference partition accounts for the flow between all the OD pairs and is appropriately expanded using the actual OD matrix using non-zero penetration rates.

A trivial solution for this limitation is to use a partition with fewer clusters that yields non zero penetration rates. In the present context, the reference partition of 5 clusters yields the full trip OD matrix as the actual OD matrix. Then, the mean penetration rates estimated from this reference partition can be assumed as reference penetration rates. For the cases of finer spatial clusters of 50, 100 and 300, since there are many OD pairs with zero trips, these so-called reference penetration rates can be used for these OD pairs to account for the flow. This technique ensures that the whole flow in the actual OD matrix is taken into account and, at the same time, penetration rates with finer spatial resolution can be obtained. Fig. 18 presents MFDs for reservoir 1 and 3 using the technique discussed. It is clear that all the results yield similar MFDs with the same peak flow and densities. Similarly, the hysteresis phenomenon is observed in all cases with a loop size similar to that of the reference case. This effect is less pronounced in the case of the

Table 5
Density of trip and real OD matrices for different levels of aggregations.

No. of clusters.	Density of trip OD matrix	Density of actual OD matrix
5 (Ref.)	1.0	1.0
50	0.165	1.0
100	0.063	1.0
300	0.011	0.99

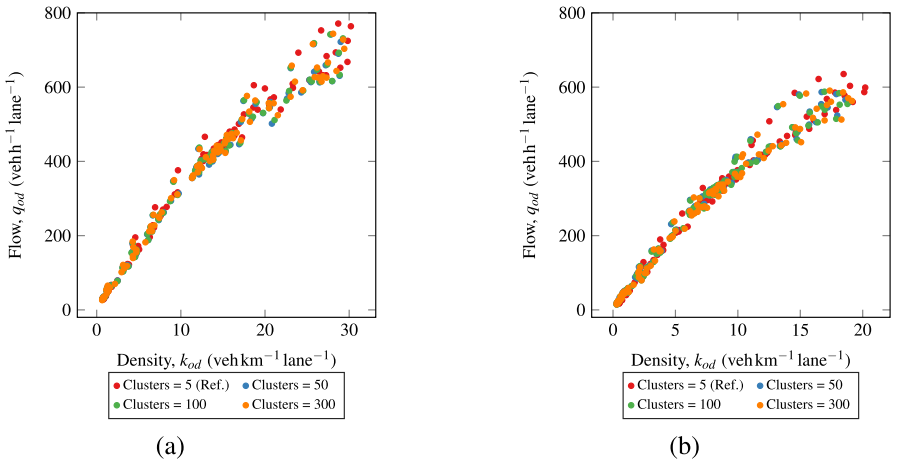

Fig. 18. Flow MFD estimates using OD specific penetration rate after correcting penetration rates. (a) Reservoir 1. (b) Reservoir 3.

origin specific penetration rate. As the origin specific penetration rate is estimated based on *all* trips that start within a given cluster, it can be observed that in the present case the penetration rate is non-zero in most of the clusters. Hence, almost all the flow from the actual OD matrix is taken into account. Fig. 19 shows MFDs for reservoirs 1 and 3 for different levels of aggregation using an origin specific penetration rate. It is evident from the plots that all the MFDs have similar characteristics and shapes irrespective of the number of clusters used. It is also clear that all the results conform with the reference partition case. Therefore, it can be concluded that there is no need to consider a refined partition to estimate the penetration rates and partitioning at regional level is sufficient. In case of using a finer spatial partition, care must be taken to define the boundaries of the partition such that the penetration rates remain non-zero.

6. Trip length estimation

6.1. Static analysis

The second part of this work presents the details of trip length estimation from mobile phone data. As stated earlier, estimating trip lengths is impossible without massive individual trajectories. Most of the works in the literature are based on network exploration methods to build virtual trips (see (Batista et al., 2019) for review). The present type of mobile phone data, *i.e.*, LBS data, provides a unique opportunity to fill the gap in trip length estimation. Since the trajectory of each trip is readily available, it is possible to represent trips based on reservoir sequences. Once the trips are clustered, major macro-paths between the macroscopic OD pairs can be identified. Note that in the present section, all the macro-paths are represented as reservoir sequences. For instance, a trip that starts in reservoir 1 and ends in 3 by crossing through reservoir 2 is indicated as $1 \rightarrow 2 \rightarrow 3$.

Let us consider the macroscopic OD pair $4 - 2$. From Fig. 5a, it is clear that there can be several possible macro-paths between the OD pair considered like $4 \rightarrow 1 \rightarrow 2$, $4 \rightarrow 3 \rightarrow 2$, *etc*. Note that these are only a few sample macro-paths and still more realistic combinations like $4 \rightarrow 3 \rightarrow 1 \rightarrow 2$, $4 \rightarrow 1 \rightarrow 3 \rightarrow 2$, *etc.* are possible in the partition considered. Assuming a single macro-path per OD

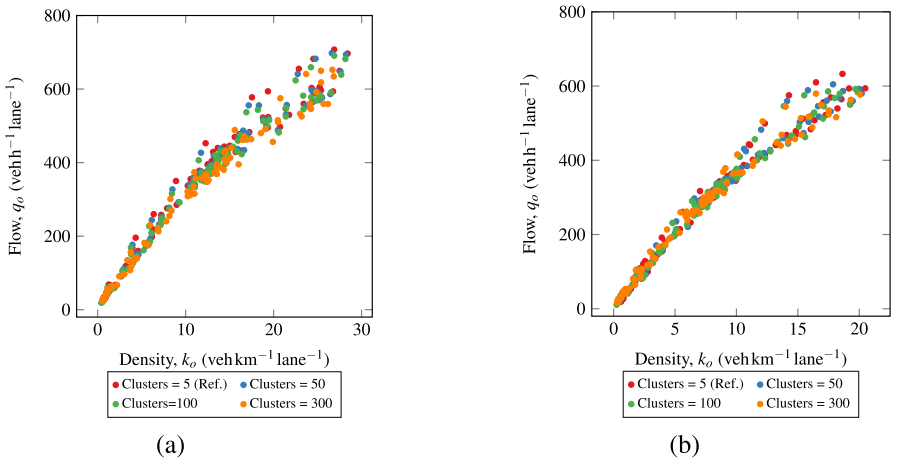

Fig. 19. Flow MFD estimates using origin specific penetration rate. (a) Reservoir 1. (b) Reservoir 3.

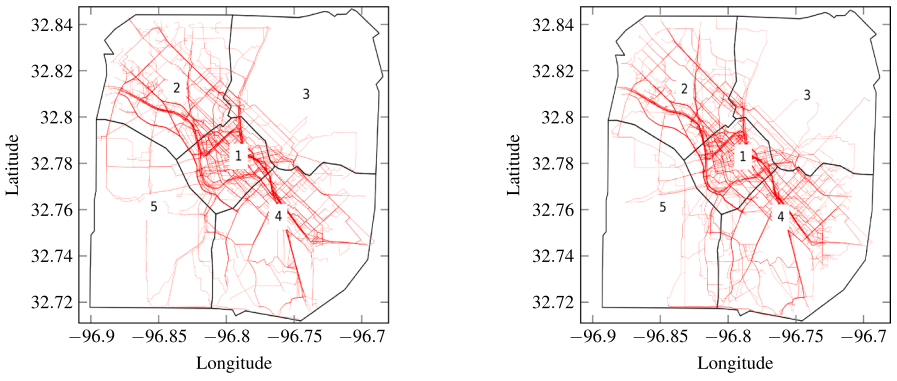

Fig. 20. Randomly sampled set of trajectories. (a) OD pair 4 − 2. (b) OD pair 2 − 4.

pair is too crude for accurately resolving traffic dynamics. On the other hand, considering all the possible and realistic macro-paths between the OD pair adds complexity to the MFD-based simulation, thereby losing the spirit of the framework. This issue can be appropriately addressed by using the trajectory data of mobile phones.

Fig. 20a shows randomly sampled trajectories estimated from phone data between the OD pair 4 − 2. At first, it is clear that most of the trips cross the network in sequence 4 → 1 → 2. Hence, it can be assumed to be the major macro-path for the OD pair considered. However, it is clear from the plot that there are other macro-paths like 4 → 3 → 1 → 2, 4 → 1 → 3 → 2, 4 → 5 → 1 → 2 and 4 → 1 → 5 → 2, which contribute towards the OD flow. The trajectory data for all weekdays can be used here to rank the most used to the least used macro-path. Consequently, macro-paths with very few trips can be safely neglected without compromising the modeling framework. In the present example of OD pair 4 − 2, almost 70% of the trips take macro-path 4 → 1 → 2 and the majority of the rest are distributed among 4 → 3 → 1 → 2 and 4 → 1 → 3 → 2. Fig. 20b presents the sampled trajectories for the OD pair 2 − 4, which is symmetrically opposite to 4 − 2. It can be observed that the macroscopic behavior of the trips is similar to that of the OD pair 4 − 2. This can be verified by the trajectory data, where the major macro-path is 2 → 1 → 4 and the rest of the flow is observed in 2 → 1 → 3 → 4 and 2 → 3 → 1 → 4. It is inferred from the data that the present network of Dallas shows this symmetric property for most of the OD pairs. However, it is a network specific property and cannot be regarded as a universal law. It should be noted that whatever the macro-path patterns are, LBS data provide enough samples to define the major macro-paths. This might not be the case with other data sources, which leads to more uncertainties.

Once the macro-paths between OD pairs are established, the following question to be solved is that of estimating the lengths of each macro-path. A straightforward and simple method, like method 1, is to estimate the average trip length inside each reservoir, taking all the trajectories into account. This corresponds to the original approach proposed by Daganzo (2007). This can be considered as constant static trip lengths, as dynamic changes in the trip lengths are neglected. This method does not account for the origin and destination of the macro-path. For instance, if the average trip length in the reservoir 1 is 1000 m, the same value is assigned for reservoir 1 for the macro-path 1 → 2 and 4 → 3 → 1 → 2. However, this can introduce big discrepancies in the MFD simulation results (Batista et al., 2019). Instead of taking only available trips from the data to estimate average trip lengths, penetration rates are also considered. For the sake of simplicity, we consider there are m number of trips for macro-path $a \to b \to c$, which are denoted as $\{L_{abc}^{a,1}, L_{abc}^{b,1}, L_{abc}^{c,1}\}$, $\{L_{abc}^{a,2}, L_{abc}^{b,2}, L_{abc}^{c,2}\}$ and so on. The notation $L_{abc}^{i,j}$ gives the length of trip number j of macro-path $a \to b \to c$ in reservoir $i \in \{a, b, c\}$. Similarly, there are n number of trips for macro-path $d \to b \to f$. According to the current framework, the penetration rates are time-dependent. For instance, depending on the departure times of m trips for macro-path $a \to b \to c$, the OD specific penetration rates for OD pair $a - c$ are denoted as $\{\rho_{ac}^1, \rho_{ac}^2, ..., \rho_{ac}^m\}$. Similarly, the origin specific penetration rates of reservoir a are represented by $\{\rho_a^1, \rho_a^2, ..., \rho_a^m\}$. A similar notation is adopted for macro-path $d \to b \to f$. Now, the mean trip length in reservoir b can be estimated using the following expressions:

$$\bar{L}_b^{od} = \frac{\sum_{j=1}^{m} \frac{L_{abc}^{b,j}}{\rho_{ac}^j} + \sum_{k=1}^{n} \frac{L_{dbf}^{b,k}}{\rho_{df}^k}}{\sum_{j=1}^{m} \frac{1}{\rho_{ac}^j} + \sum_{k=1}^{n} \frac{1}{\rho_{df}^k}}, \quad \text{and} \quad \bar{L}_b^{o} = \frac{\sum_{j=1}^{m} \frac{L_{abc}^{b,j}}{\rho_{a}^j} + \sum_{k=1}^{n} \frac{L_{dbf}^{b,k}}{\rho_{d}^k}}{\sum_{j=1}^{m} \frac{1}{\rho_{a}^j} + \sum_{k=1}^{n} \frac{1}{\rho_{d}^k}}, \tag{7}$$

where $\bar{L}_b^{od}$ is the mean trip length based on the OD specific penetration rate and $\bar{L}_b^{o}$ is the mean trip length using the origin specific penetration rate. Thus, the mean trip length inside a reservoir can be interpreted as the weighted mean of all the trips crossing this reservoir using the inverse of either the OD or origin specific penetration rates as weights.

Table 6 presents average trip lengths along with standard deviations inside each reservoir estimated using OD and origin specific penetration rates. It can be seen that the average trip lengths are very similar using both types of penetration rate. The lengths are representative of the size of the reservoirs considered. The standard deviation values are relatively large, indicating that there is huge variation of trip lengths inside each reservoir. A more intensive method, like method 2, is to average trip lengths inside each reservoir based on the macro-path, *i.e.*, taking the mean of all the trips that cross a given reservoir for a given macro-path. Again, this approach

Table 6
Average trip length and standard deviations in m inside each reservoir using OD and origin specific penetration rates using method 1.

Reservoir	Average trip length using OD pen. rate (m)	Average trip length using origin pen. rate (m)
1	1959 ± 1440	1921 ± 1450
2	2186 ± 1867	2035 ± 1766
3	2132 ± 1872	2090 ± 1811
4	2593 ± 2106	2484 ± 2116
5	2751 ± 2026	2494 ± 1978

is possible only because of the massive phone data available, which guarantees enough observations on all the major macro-paths. The difference between the two approaches is that the former considers the average of all the trips inside each reservoir irrespective of the macro-path, while the latter considers the mean trip length inside each reservoir for a given macro-path individually. In other words, the average trip length in the reservoir, for example 1, is the same in all the macro-paths in method 1. Whereas in method 2, the mean trip length of reservoir 1 changes for each macro-path. Batista et al. (2019) stated that method 2 in the present context is more accurate than method 1, based on the simulation results. In the current work, discrepancies between the two methods are demonstrated using phone data.

Table 7 shows the average trip lengths along with the standard deviations of selected macro-paths using both approaches and the relative differences in total trip lengths. Consider the internal macro-path $1 \rightarrow 1$. The mean trip length estimated by method 1 is 1959 m, while method 2 gives 1370 m, which is significantly lower than its counterpart. As reservoir 1 is in the downtown area of Dallas, the majority of internal trips are between the freeways that encompass the area. One of the longest trips possible, without considering the freeway network, in this reservoir is around 2500 m. Taking this into account, an average trip length of 1959 m over more than 100000 trips is unrealistic. The reason for such a high mean trip length is due to averaging all the trips that cross reservoir 1 irrespective of OD pair. This can be elaborated clearly using macro-paths $3 \rightarrow 1 \rightarrow 5$ and $4 \rightarrow 1 \rightarrow 2$. In Fig. 5a, it is clear that both stated macro-paths need to cross the reservoir in its entirety. As they are mostly long distance trips, users tend to use freeways, which are ring roads in reservoir 1. Hence, longer average trip lengths are observed for these macro-paths in reservoir 1, as vehicles need to circumnavigate the downtown area. In method 1, these types of trip are aggregated along with the internal trips of reservoir 1, so a higher average trip length is estimated. On the other hand, method 2 estimates a more representative shorter trip length for internal trip $1 \rightarrow 1$ and a longer trip length for macro-paths $3 \rightarrow 1 \rightarrow 5$ and $4 \rightarrow 1 \rightarrow 2$, as expected. This conclusion is in-line with the results in the literature (Batista et al., 2019) and the present findings demonstrate that phone data provides a practical and effective way to calibrate the trip lengths.

The standard deviation of the trip lengths is estimated inside each reservoir to characterize the diversity of estimated lengths that fall within the same macro-path. Note that the estimation of the standard deviation of the macro-paths is also very high for the DTA framework (Batista and Leclercq, 2019). Table 8 presents the trip lengths for the selected macro-paths using method 2 and their corresponding standard deviations. In the case of internal trips, *i.e.*, $1 \rightarrow 1$ and $2 \rightarrow 2$, the standard deviation is quite large compared to the mean trip length. This is expected as trip lengths can vary widely inside the reservoir and hence there is a higher coefficient of variation. However, for macro-paths that cross reservoir 1, for instance $2 \rightarrow 1 \rightarrow 5$, $4 \rightarrow 1 \rightarrow 5$, *etc*., the coefficient of variation inside reservoir 1 is relatively low. This shows that the estimated macro-paths effectively captured the trip patterns; the average trip length across reservoir 1 is very similar for all the trips. The coefficient of variation in the origin and destination reservoirs for the stated macro-paths is relatively large. This is due to the fact that the exact points of departure and arrival can vary across a wide range in the reservoirs and hence lead to a wide range of trip lengths. It is also noteworthy that the partition results considered in reservoirs that are relatively large in area and using a finer partition will decrease the variability of trip lengths.

In order to illustrate the importance of considering average trip length per macro-path, we consider two macro-paths $4 \rightarrow 3 \rightarrow 1 \rightarrow 2$ and $4 \rightarrow 1 \rightarrow 3 \rightarrow 2$. Although the same reservoirs are crossed, albeit in a different order, between the OD pair considered, the mean trip lengths inside each reservoir for both trips differ from each other. Figs. 21 and 22 show the trip length distributions for the two stated macro-paths, respectively. It is clear from the distributions that even for the same OD pair, the mean trip lengths inside each reservoir depend on the macro-path. In the trip length distributions shown in the plots, mean trip lengths show significant differences except for mean trip length inside reservoir 2. For instance, the mean trip length in reservoir 1 for macro-path $4 \rightarrow 1 \rightarrow 3 \rightarrow 2$ is 3461 m, whereas in the case of $4 \rightarrow 3 \rightarrow 1 \rightarrow 2$, it increases to 4079 m. Hence, this inference reinforces the previous conclusion on the importance of considering mean trip lengths per OD pair and per macro-path.

Table 7
Average trip lengths per reservoir and per macro-path and standard deviations in m using the OD specific penetration rate for selected macro-paths.

Macro-path	Avg. length per reservoir (m)	Avg. length per macro-path (m)	Rel. difference in total length (%)
$1 \rightarrow 1$	{1959 ± 1440}	{1370 ± 1063}	43
$1 \rightarrow 3$	{1957 ± 1440, 2132 ± 1872}	{1736 ± 1336, 2341 ± 1986}	0.3
$3 \rightarrow 1 \rightarrow 5$	{2132 ± 1872, 1959 ± 1440, 2751 ± 2026}	{2817 ± 1944, 4582 ± 863, 3502 ± 1968}	37
$4 \rightarrow 1 \rightarrow 2$	{2593 ± 2106, 1959 ± 1440, 2186 ± 1867}	{2959 ± 2070, 4549 ± 925, 3129 ± 1914}	36

Table 8
Average trip lengths and standard deviations in m using the OD specific penetration rate for selected macro-paths.

Macro-path	Avg. length per reservoir (m)	Coefficient of variability
1 → 1	{1370 ± 1063}	{0.77}
2 → 2	{1758 ± 1505}	{0.85}
2 → 1 → 5	{2898 ± 1948, 2748 ± 907, 3194 ± 1915}	{0.67, 0.32, 0.59}
4 → 1 → 5	{2911 ± 2068, 3859 ± 1229, 3317 ± 1975}	{0.71, 0.31, 0.59}
4 → 3 → 1 → 2	{2127 ± 1546, 2196 ± 1240, 4079 ± 1108, 2793 ± 1835}	{0.72, 0.56, 0.27, 0.65}
5 → 1 → 4	{2957 ± 1780, 3725 ± 1349, 3338 ± 2089}	{0.60, 0.36, 0.62}

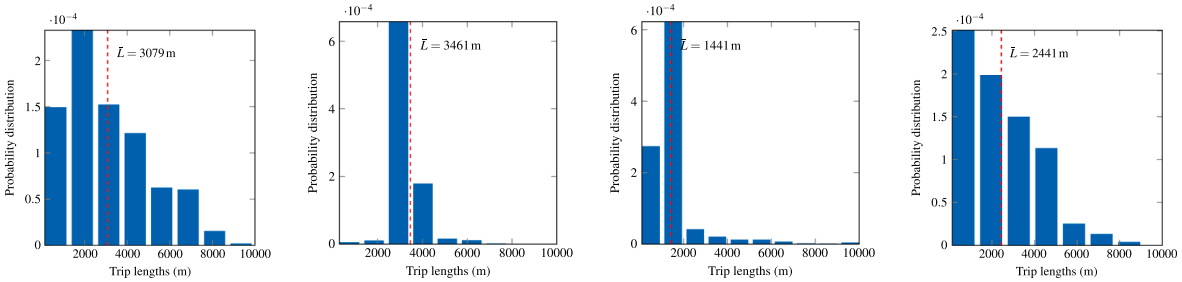

Fig. 21. Distribution of trip lengths inside each reservoir for the macro-path 4 → 1 → 3 → 2. (a) Reservoir 4. (b) Reservoir 1. (c) Reservoir 3. (d) Reservoir 2.

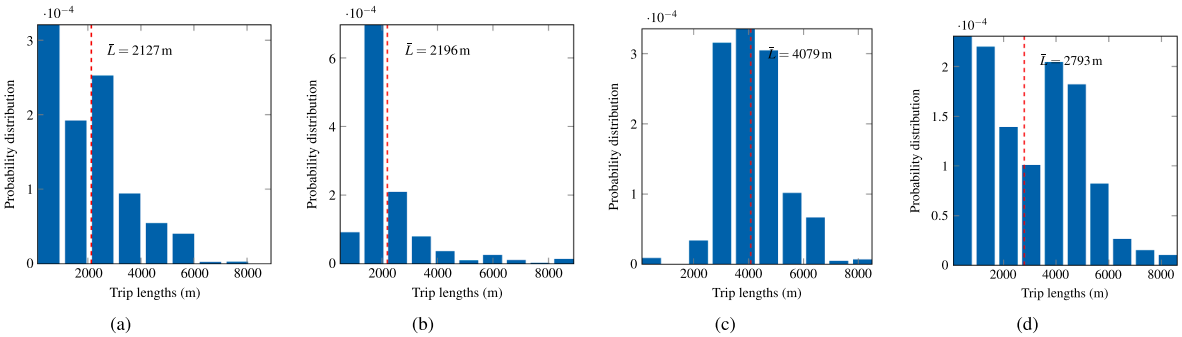

Fig. 22. Distribution of trip lengths inside each reservoir for macro-path 4 → 3 → 1 → 2. (a) Reservoir 4. (b) Reservoir 3. (c) Reservoir 1. (d) Reservoir 2.

6.2. Dynamic analysis

The analysis presented until now has neglected the dynamic information regarding trip lengths, *i.e.*, changes in the trip lengths with the time of the day and the traffic conditions. Only a few works appear to have dealt with the dynamic variation of trip lengths using empirical data due to a lack of data. Local congestion can influence trip length distributions, as users tend to avoid shorter congested routes to take longer ones. The importance of considering the variation of mean trip lengths with time was discussed in (Yildirimoglu and Geroliminis, 2014; Yildirimoglu et al., 2018) in the context of simulation studies. The average trip lengths presented in Tables 6, 7 and distributions in Figs. 21, 22 take into account all the trips observed during the whole 2-month period. However, it is intuitive that trip lengths within and between reservoirs tend to be dynamic, as the users tend to prefer alternative paths during congestion periods. Batista et al. (2020) proposed a framework to estimate dynamic trip lengths explicitly and concluded that including dynamic variations in trip lengths improves the accuracy of MFD-based simulations. However, the authors of this work built a virtual set of trips using a network exploration technique to perform this analysis rather than empirical trip lengths.

In the present work, dynamic trip lengths between macroscopic OD pairs are estimated using the departure time of each trip. To do so, an aggregate time period of 60 min is considered and hence for each day there are 24 aggregation periods. For a given OD pair, all the trips that start within a given aggregation period are selected and a mean trip length is estimated for that given period. This is done for each weekday separately in the 2 month period considered. Finally, the mean trip length within each aggregation interval is estimated for all weekdays. Fig. 23 presents the mean evolution of trip lengths along with confidence intervals at each aggregate interval for different macroscopic OD pairs. The confidence intervals are estimated for 95% of confidence level assuming a normal distribution. The macro-paths that show different types of trends are selected for the discussion.

Consider Fig. 23a, where the evolution of internal macro-path 3 → 3 is presented. Two peaks, one at morning peak hour and

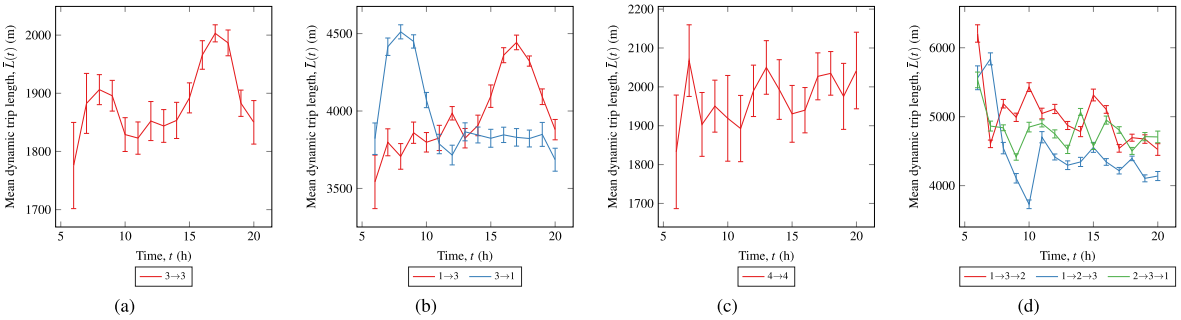

Fig. 23. Mean dynamic trip lengths and their standard deviations for different macro-paths. (a) Internal macro-path 3 → 3. (b) Macro-paths 1 → 3 and 3 → 1. (c) Internal macro-path 4 → 4. (d) Macro-paths crossing reservoirs 1, 2 and 3.

another at evening peak hour, can be clearly seen in the plot. This trend signifies that the users perhaps take longer paths during peak hours to avoid the most congested routes. The difference between trip lengths at the peak and non-peak periods is less than 10% for macro-path 3 → 3. However, this relatively insignificant difference can introduce a considerable bias in the traffic dynamics in MFD-based simulations (Batista et al., 2019). Fig. 23b shows the dynamic trip lengths of the macro-paths 1 → 3 and 3 → 1. It is clear from the plots that the dynamic trip length evolutions of two macro-paths are nearly symmetric. This is due to the movement of people from suburban regions to downtown in the morning and *vice versa* in the evening. The evolution of the mean dynamic trip length of the macro-path 4 → 4 is shown in Fig. 23c, where the changes in trip length are relatively insignificant. This is due to the presence of large urban spaces for leisure activities, which decreases its overall contribution to the flow of the network. This can be justified using the total network length presented in Table 4, where reservoir 4 has the shortest network length of all the reservoirs. Hence, a more stable trip length evolution is observed. Finally, Fig. 23d presents the dynamic trip lengths of the macro-paths that involve reservoirs 1, 2 and 3 in different sequences. Although all the macro-paths present variations within the same limits of trip lengths, no clear trend in evolution is observed. These types of behaviors are justified, given the topology of the network. However, it is difficult to predict these trends between different OD pairs *a priori* and appropriately calibrate the MFD models. It is also clear from the plots that the confidence intervals remain relatively small indicating reliable estimation of trip lengths. The confidence intervals for the first aggregation period tends to be large in some cases. This can be due to the fewer data available during early hours in the morning, which adds noise to the trip length estimation. This type of analysis has not been performed before in the literature due to the lack of sufficient and reliable data. Therefore, to the knowledge of the authors, the work presented here is the first to estimate the dynamic evolution of the trip lengths of different macro-paths empirically, making it possible to used it directly in the MFD simulation framework.

6.3. Empirical relation between trips lengths and mean speed

It is clear from Figs. 23a and b that the mean trip lengths change according to traffic conditions and tend to increase during congestion periods. Based on this inference, existence of well-defined relationship between trip lengths and mean speed is investigated. Unraveling such a relationship is of particular interest for the MFD-based models as it provides a direct correlation to adjust the trip lengths with the network traffic states. As trip lengths and mean speeds can vary widely depending on macro-path and OD pair, these quantities must be normalized to establish a relationship. To this extent, detour ratio is chosen as the surrogate to trip lengths. Detour ratio is assumed as the ratio of the actual trip distance to the Haversine distance (great circle distance) between origin and destination of a trip. Similarly, mean speeds are normalized by free-flow speed, which is assumed as the 90th percentile of all observed mean speeds for a given OD pair. It is to be noted that the free-flow speed is defined per OD pair and consequently, mean speeds of a given OD pair are normalized with the free-flow speed of that OD pair.

Fig. 24 presents the relationship between the mean detour ratio and the normalized mean speeds. In order to compute the mean detour ratio, normalized mean speeds are divided into several bins and the mean detour ratio inside each bin is considered. The plots include 95% confidence intervals to show the robustness of empirical data. Fig. 24a shows the trend for the entire network and Fig. 24b–f present the relationship for each reservoir. First and foremost, well-defined relationships are obtained for all reservoirs. The trend is in agreement with previous conclusions that at lower mean speeds, users tend to take detours to avoid heavily congested routes and thus, higher detour ratios are observed. Note that for very low mean speeds, the detour ratio seems to decrease. This is certainly because of heavy congestion spreads over the full network eventually, making detour useless. A higher mean speed over a trip indicates shorter path in distance and hence, lower detour ratio. It is also clear that the confidence intervals are relatively small, which implies reliable estimation. A non-linear decaying functional form can be used to fit empirical data and an exponential functional form is considered in the current work as follows,

$$d_R = a + b\, e^{c\,\widetilde{v}}, \tag{8}$$

where d_R is detour ratio and $\widetilde{v}$ is normalized mean speed.

Table 9 presents fit coefficients and corresponding R^2 values for each reservoir and entire network. It is clear from the fit coefficients that all reservoirs except reservoir 2 and the entire network have very similar fit characteristics and tight fits. This is very

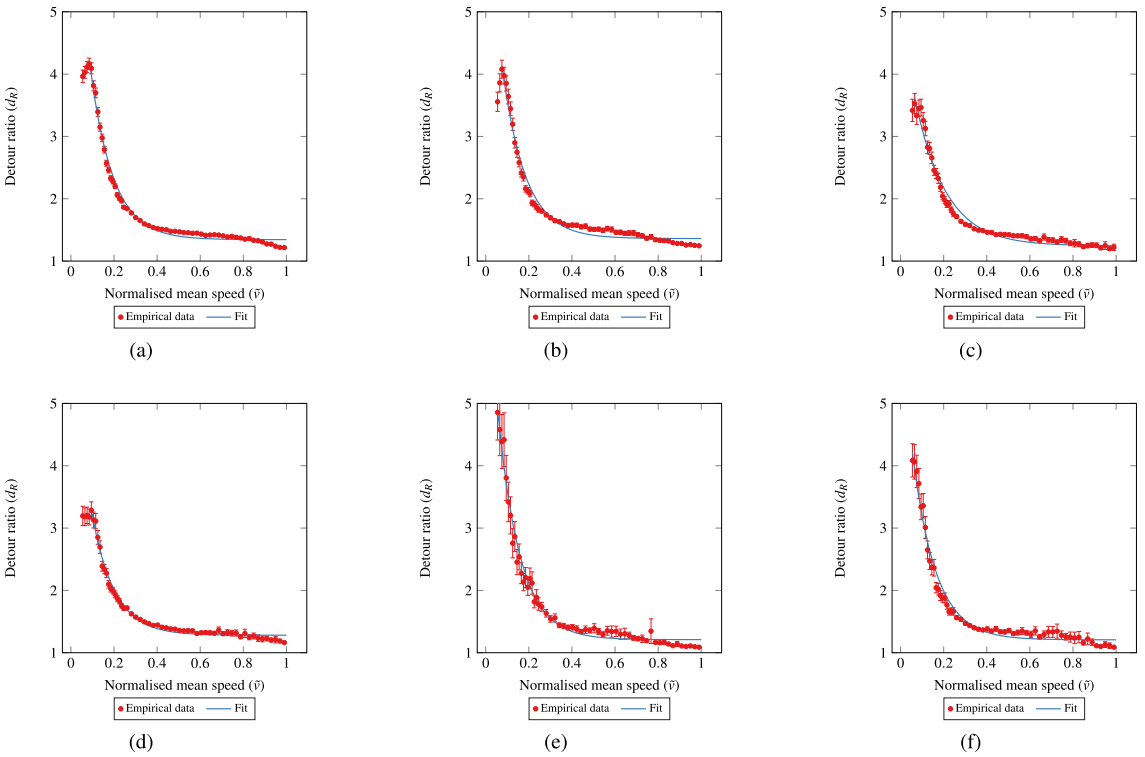

Fig. 24. Relationship between detour ratio and normalized mean speeds. (a) Entire network. (b) Reservoir 1. (c) Reservoir 2. (d) Reservoir 3. (d) Reservoir 4. (e).

Table 9
Fit parameters and R^2 values for mean speed - detour ratio relationship.

Reservoir	Fit parameters			R^2
	a	b	c	
All	1.34	6.44	−9.47	0.99
1	1.36	5.57	−9.29	0.98
2	1.24	3.74	−6.96	0.97
3	1.28	5.01	−9.64	0.98
4	1.21	6.26	−9.68	0.98
5	1.21	5.01	−9.29	0.98

important result as it shows that there is no need for specific analysis at the regional level to characterize the trip lengths dependency to the traffic conditions. Further research must be done with data from different network topologies to verify the existence of such relationship and its universality. However, this first step looks promising because of very similar fit coefficients observed in the present case. Furthermore, this analysis show that the detour ratio at high (free-flow) mean speeds falls between range 1.2 to 1.3, see coefficient a in Table 9. This confirms the previous observations from the literature (see Yang et al. (2018a) and references within). As only straight line distance between the origin and the destination of a trip is used to estimate the detour ratio, this relationship can be used for the trajectories with poor spatial resolution to estimate the actual trip distance. This is also remarkable finding as this can be used to provide a rough estimate of trip lengths without requiring extensive analysis of experimental data or the network topology.

7. Estimation of path flow distribution

The principal input data for any MFD-based simulation are the underlying MFD, the macro-paths and their corresponding trip lengths, which have been dealt with so far. As seen earlier, there can be more than one feasible macro-path between a given macroscopic OD pair. It is clear that almost all the major macro-paths, *i.e.*, those that have higher flows compared to the others in the present partitioning, cross reservoir 1, which is the downtown area. Thus, considering just one major macro-path between an OD pair and assigning the total flow to this path might lead to unrealistically high flows in reservoir 1, which might result in a gridlock. Hence, depending on the relative flow between all the feasible macro-paths for a given OD pair, it is necessary to have more than one macro-path. For instance, for internal trips that start and finish in reservoir 1, it can be observed that 97% of the trips have the macro-

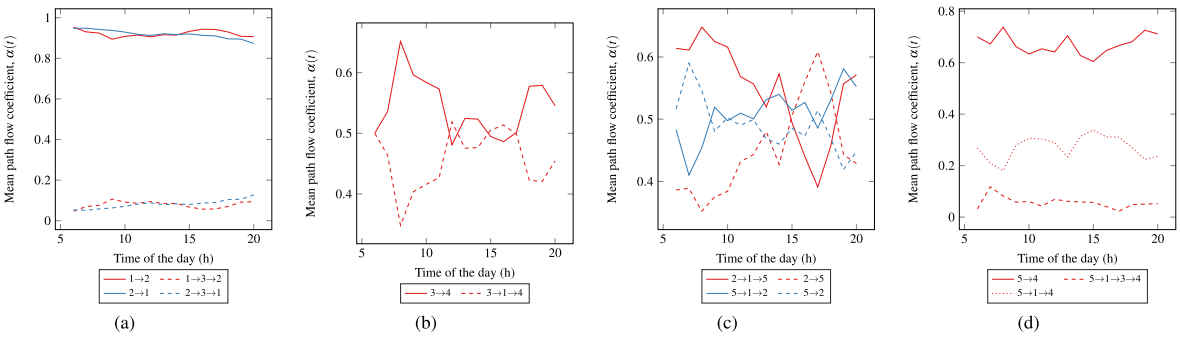

Fig. 25. Mean evolution of path flow coefficients between different macroscopic OD pairs. (a) OD pairs $1-2$ and $2-1$. (b) OD pair $3-4$. (c) OD pairs $2-5$ and $5-2$. (d) OD pair $5-4$.

path $1 \rightarrow 1$ and the rest follows $1 \rightarrow 2 \rightarrow 1$, $1 \rightarrow 3 \rightarrow 1$, *etc*. In this case, it is safe to neglect the other macro-paths and consider only the major macro-path. It can be observed that all the internal trips in the present work show this behavior, so only one major macro-path is considered for them. However, this is far from true for transfer trips that have origins and destinations in different reservoirs.

In the case of having more than one macro-path between an OD pair, it is essential to know the amount of flow to be assigned to each path. This shows how far the observed network equilibrium is from UE conditions and shed some light about the validity of UE principle at large scale. In this context, the path flow coefficient for a macro-path p for a given OD pair, od can be defined as:

$$\alpha_{od}^{p} = \frac{N_{od}^{p}}{\sum_{i \in \mathscr{P}_{od}} N_{od}^{i}}, \tag{9}$$

where N_{od}^{i} is the number of trajectories on macro-path i and $\mathscr{P}_{od}$ is the set of all the major macro-paths between the OD pair, od. The path flow coefficient can be estimated using DTA determining the UE conditions. However, DTA can be computationally demanding depending on the size of the network under study, and it is possible to extract the information on path flow coefficients using trajectory data. However, at the network level, it is not possible to observe the path flow distribution between all local OD pairs. This becomes feasible only at the regional level and is extracted using phone data. Since path flow distributions are computed assuming UE settings, it is possible to validate this hypothesis by empirically deriving the gap with the UE conditions from phone data. The important research questions in this context are to determine: (i) if the gaps for macroscopic OD pairs are close to zero, as assumed by the UE hypothesis in DTA, and (ii) if the gaps change in time.

Fig. 25 shows the mean evolution of the path flow coefficients for a few OD pairs. There are two major macro-paths each between the OD pair $1-2$ and $2-1$ and their corresponding path flow evolutions are presented in Fig. 25a. It can be observed that macro-paths $1 \rightarrow 2$ and $2 \rightarrow 1$ have the majority of the flow in both cases and they remain stable over the course of the day. On the other hand, the path flow for the macro-paths between OD pair $3-4$ shows a significant variation, which is shown in Fig. 25b. It is clear that during peak hours, users tend to use macro-path $3 \rightarrow 4$ over $3 \rightarrow 1 \rightarrow 4$, whereas both macro-paths undergo nearly equal amounts of flow during the off peak hours. Fig. 25c shows a similar trend, although users tend to take one macro-path during the morning peak and another during the evening peak. Consider the OD pair $2-5$, where the users prefer macro-path $2 \rightarrow 1 \rightarrow 5$ over the macro-path $2 \rightarrow 5$ during the morning peak hour and *vice versa* during the evening peak hour. A symmetrically opposite case is observed in the reverse trip, *i.e.*, for OD pair $5-2$, where users use the macro-path $5 \rightarrow 2$ in the morning over the macro-path $5 \rightarrow 1 \rightarrow 2$. Finally, Fig. 25d shows the path flow coefficients of the three major macro-paths for OD pair $5-4$, where a stable evolution in all three macro-paths is observed. With such valuable information, the DTA step in MFD-based models can be eliminated. The models can be fed directly with the dynamic variation of trip lengths and path flow distributions. Since these parameters are directly estimated from the empirical phone data, the resulting traffic dynamics from MFD-based simulations will be closer to real data. At the same time, estimating parameters like dynamic trip lengths and path flow distributions is very difficult with other types of data source like LDD.

8. Estimation of user equilibrium gap

Finally, the current work is concluded by presenting the estimates of UE gaps for each macroscopic OD pair. The gap corresponds to the relative difference between the travel time on a given macro-path and the minimum travel time among all the macro-paths for a given OD pair (Sbayti et al., 2007). The UE gap for a given OD pair, G_{od} can be expressed as:

$$G_{od} = \frac{1}{TT_{min}^{od}} \sum_{i \in \mathscr{P}_{od}} \alpha_{od}^{i} (TT_{i}^{od} - TT_{min}^{od}), \tag{10}$$

where TT_{i}^{od} is the travel time of macro-path i, TT_{min}^{od} is the minimum travel time between the OD pair od and α_{od}^{i} is the path flow coefficient of the macro-path. If all the macro-paths between a given OD pair are the shortest paths in time, G_{od} is zero by definition. On the other hand, if the macro-path that has the longest travel time undergoes a higher flow, the gap is bigger and the network is far from the equilibrium. Hence, this parameter shows how far the network is from UE conditions. The dynamic gap evolution can be

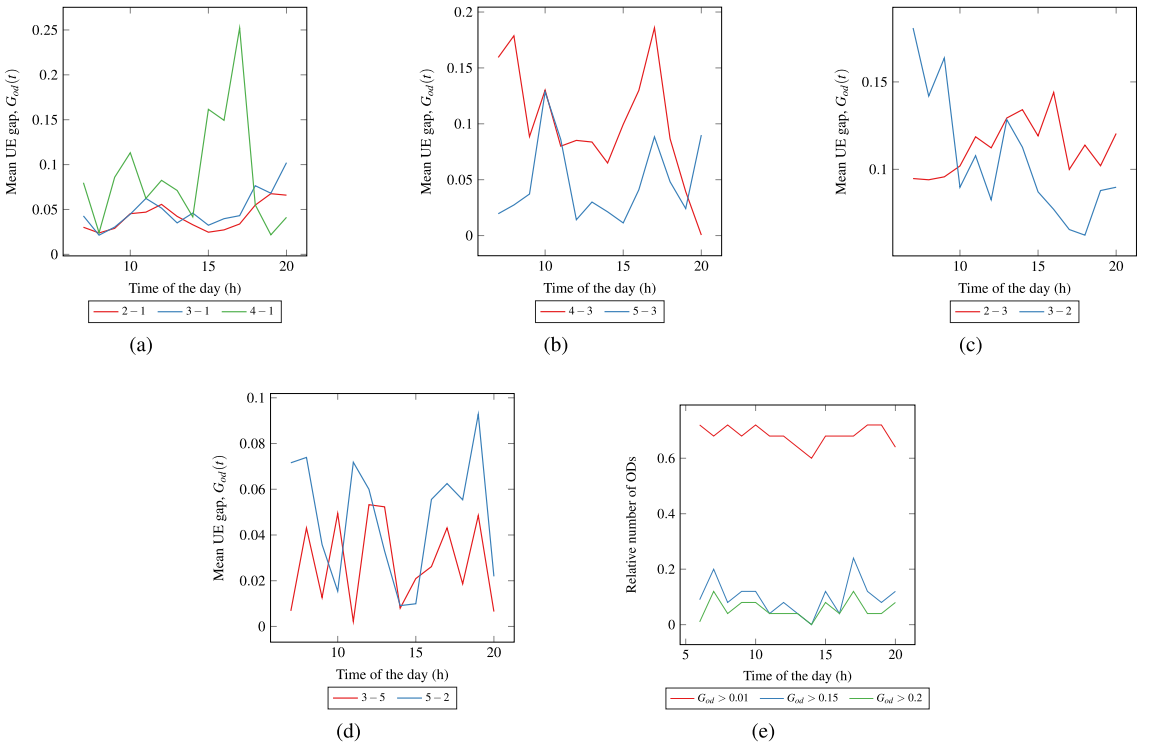

Fig. 26. Mean evolution of relative UE gap for different macroscopic OD pairs. (a) OD pairs that involve the downtown area, *i.e.*, reservoir 1. (b) OD pairs that have peak hour distributions. (c) OD pairs that have symmetric peak hours. (d) OD pairs with low mean gap. (e) Relative number of ODs for a given gap condition.

computed by using eq. (10) within each aggregation period of 60 min. The trips starting within a given aggregation period are collected for a given OD pair and the minimum travel time amongst all the macro-paths is estimated. Using this minimum travel time and path flow coefficients computed earlier, it is trivial to estimate the dynamic UE gap, $G_{od}(t)$.

Fig. 26 presents the evolution of UE gaps for different macroscopic OD pairs. The selected OD pairs show different trends as discussed in the previous cases. Fig. 26a shows the OD pairs that involve downtown Dallas and the neighboring suburbs. Since the OD pair $5-1$ has only one major macro-path, $G_{53}(t)$ is always zero. It can be observed that the OD pairs $2-1$ and $3-1$ have very low gap values, where the gaps observed are less than 10%. However, OD pair $4-1$ shows the peaks in the morning and evening with relatively high gap values. For the OD pairs $2-1$ and $3-1$, the major macro-paths are $2 \rightarrow 1$ and $3 \rightarrow 1$, respectively, where on average more than 90% of the users choose this path. In the case of OD pair $4-1$, the proportion of users choosing macro-path $4 \rightarrow 1$ is comparatively low compared to other OD pairs. More than 20% choose the longer macro-path of $4 \rightarrow 3 \rightarrow 1$ in this case. Hence, bigger UE gaps are noticed for this particular OD pair. Fig. 26b presents the gaps for OD pairs $4-3$ and $5-3$, where peaks in the

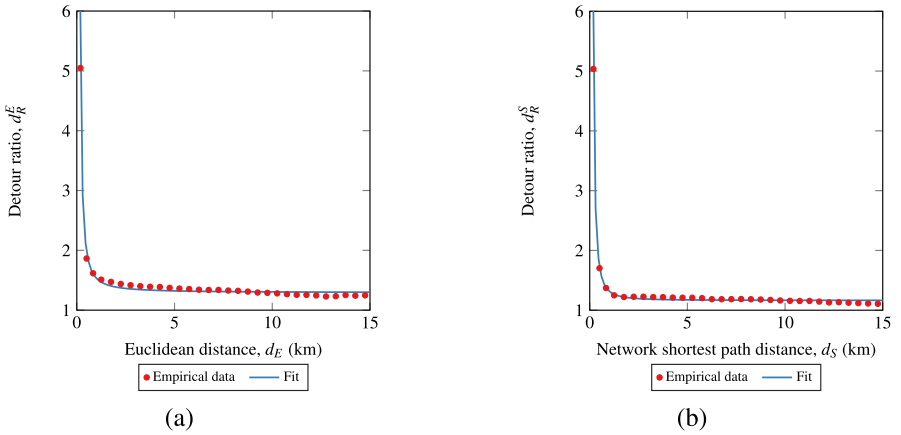

Fig. 27. Trip detour ratio *vs.* Euclidean distance in Dallas, TX. (a) Using Euclidean distance. (b) Using network shortest path distance.

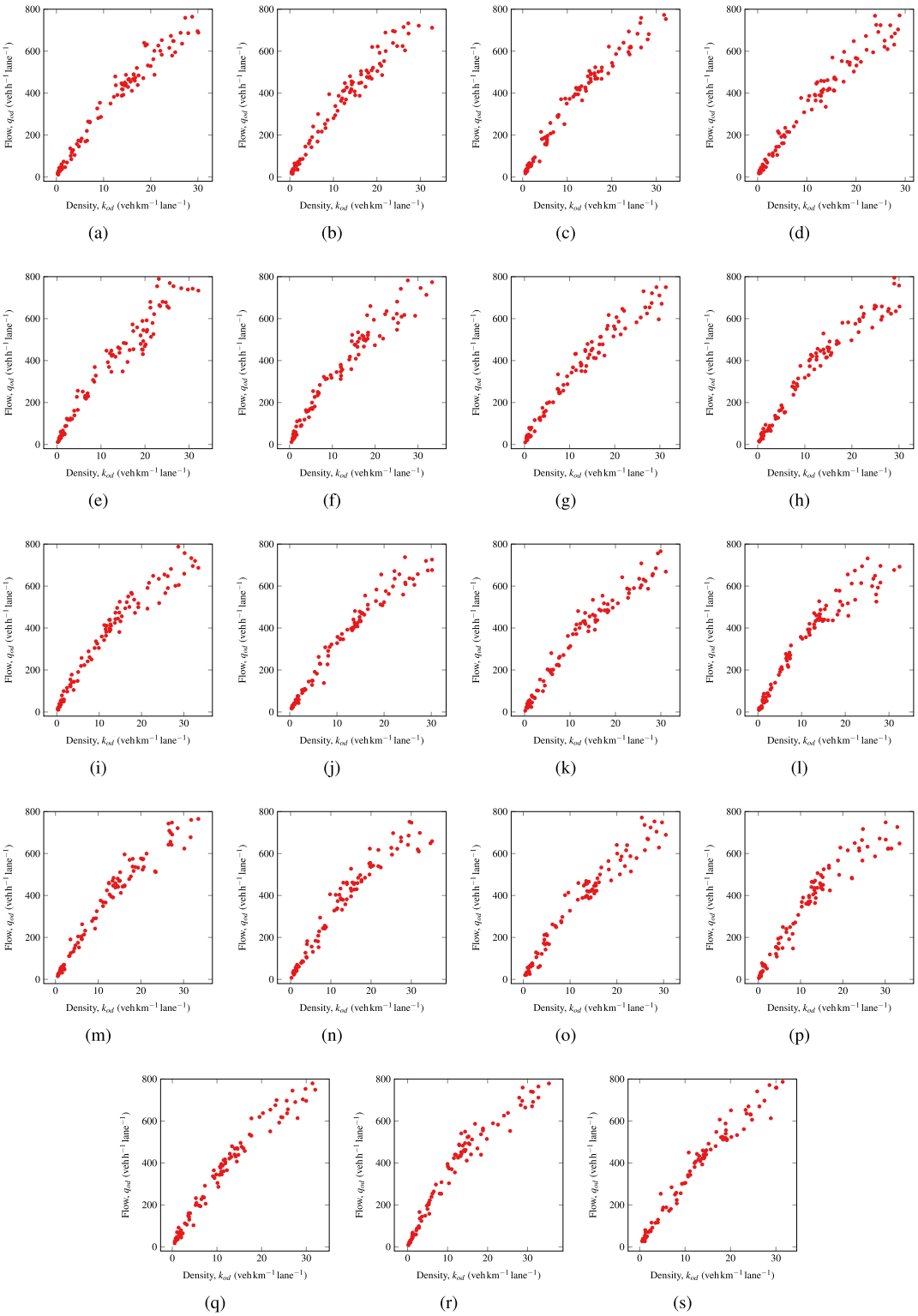

Fig. 28. Day-to-day Flow MFD estimates using the OD specific penetration rate for reservoir 1. (a) Thu 2/3/2017. (b) Fri 3/3/2017. (c) Mon 6/3/2017. (d) Tue 14/3/2017. (e) Thu 16/3/2017. (f) Mon 20/3/2017. (g) Tue 21/3/2017. (h) Wed 22/3/2017. (i) Tue 28/3/2017. (j) Wed 29/3/2017. (k) Thu 30/3/2017. (l) Mon 3/4/2017. (m) Tue 4/4/2017. (n) Fri 7/4/2017. (o) Mon 10/4/2017. (p) Tue 11/4/2017. (q) Wed 19/4/2017. (r) Wed 26/4/2017. (s) Fri 28/4/2017.

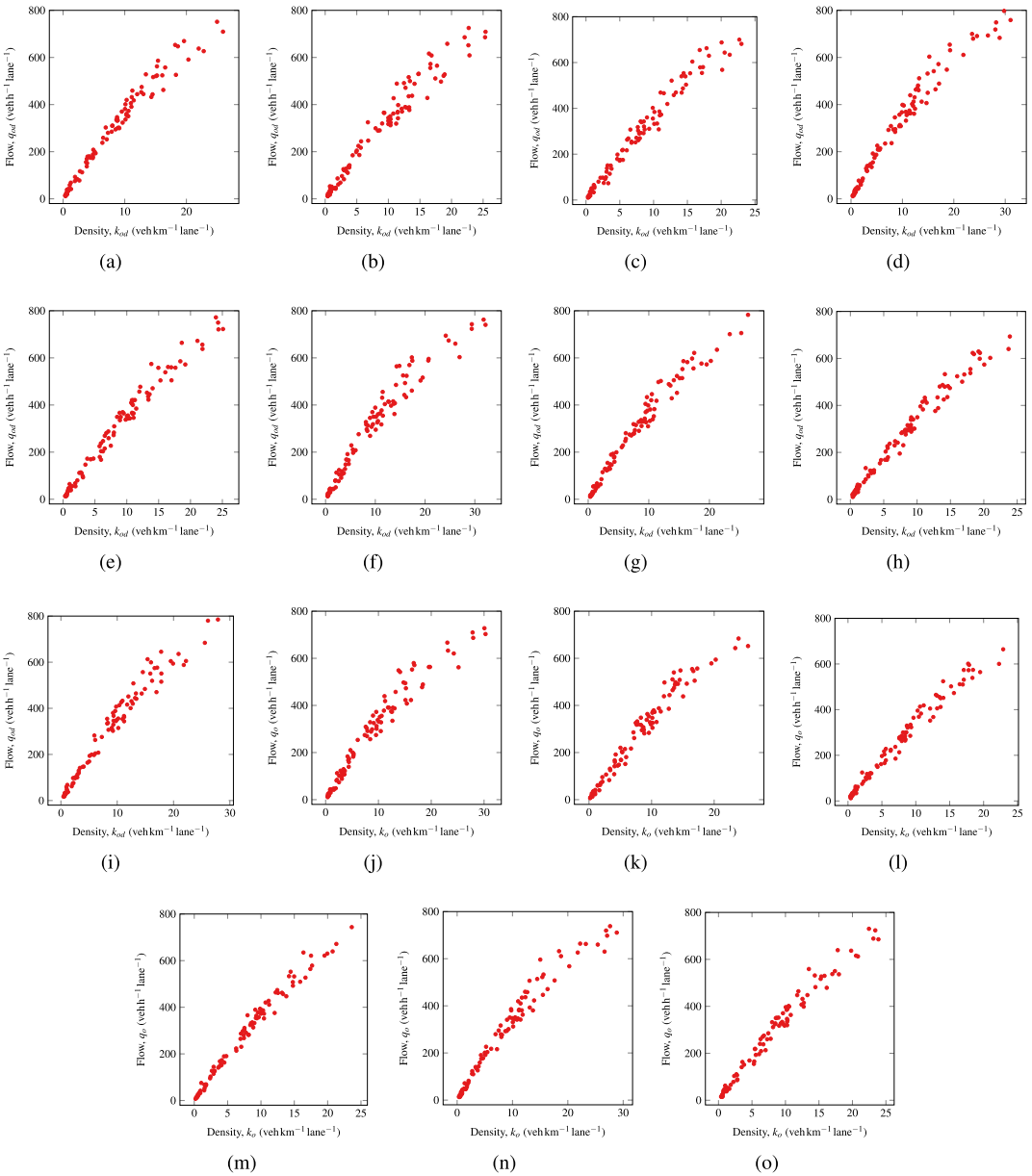

Fig. 29. Day-to-day Flow MFD estimates using the OD specific penetration rate for reservoir 2. (a) Thu 9/3/2017. (b) Fri 10/3/2017. (c) Tue 14/3/2017. (d) Wed 15/3/2017. (e) Fri 17/3/2017. (f) Wed 22/3/2017. (g) Thu 30/3/2017. (h) Fri 31/3/2017. (j) Wed 5/4/2017. (k) Fri 7/3/2017. (l) Thu 30/3/2017. (m) Mon 10/4/2017. (n) Wed 19/4/2017. (o) Thu 20/4/2017. (p) Fri 28/4/2017.

morning peak and the evening peak periods can be observed. Symmetric peak hours can be noticed for the OD pairs $2-3$ and $3-2$ in Fig. 26c, where OD pair $3-2$ has a bigger gap during the morning peak hour and its counterpart has a bigger gap during the evening peak hour. This is clearly due to the difference in the direction of traffic flow during the morning and evening periods. Fig. 26d shows the mean UD gap evolution for OD pairs $3-5$ and $5-2$, where cyclic variations are obtained. At the same time, it is also clear that the magnitude of the gap is relatively low for the stated OD pairs, where the day averaged gap is close to 5%.

Finally, Fig. 26e presents the time evolution of the relative number of ODs for a given gap condition. For instance, the red curve corresponds to the number of ODs with gap $G_{od} > 0.01$. It is clear from the plot that almost 70% of the total OD pairs have gaps exceeding 0.01. In other words, if the network is assumed to be in UE condition when the gap less than or equal to 0.01, from the plot it is evident that less than 30% of ODs fulfill the UE condition. It is trivial that if the threshold is increased, the number of ODs satisfying the UE conditions increase. This can be observed in the plot, where threshold gaps of 0.15 and 0.2 are also presented. In the cases of threshold gaps of 0.15 and 0.2, two peaks, one in the morning rush hour and another in the evening can be clearly noticed. This is noticeable as the validity of the UE principle looks less likely during peak hours, at least for few OD pairs. Overall, if the

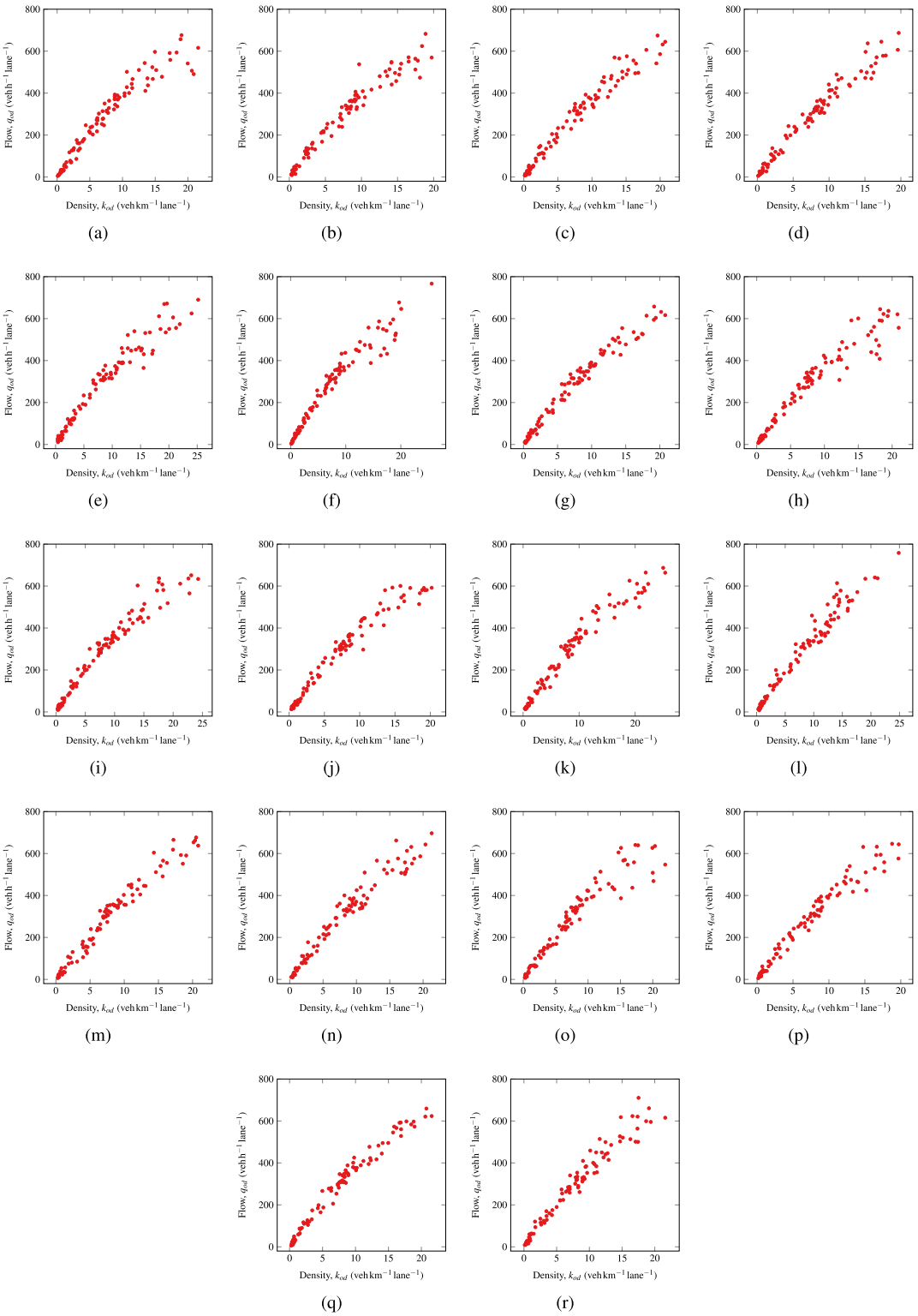

Fig. 30. Day-to-day Flow MFD estimates using the OD specific penetration rate for reservoir 3. (a) Mon 6/3/2017. (a) Mon 20/3/2017. (b) Tue 21/3/2017. (c) Wed 22/3/2017. (d) Fri 31/3/2017. (e) Wed 5/4/2017. (f) Tue 4/4/2017. (g) Fri 7/4/2017. (h) Mon 10/4/2017. (j) Thu 13/4/2017. (k) Fri 14/4/2017. (l) Thu 20/4/2017. (m) Fri 21/4/2017. (n) Mon 24/4/2017. (o) Tue 25/4/2017. (p) Wed 26/4/2017. (q) Thu 27/4/2017. (r) Fri 28/4/2017.

threshold gap of 0.15 or more is assumed, almost 90% of ODs pairs satisfy the UE conditions, which is reassuring for the classical simulations based on this principle, but at the same time it is not completely satisfactory. From a model implementation point-of-view, it is important to identify OD pairs that deviate from this principle and collect more data from such cases to understand the causes for deviation. In the current case, it is noticed that OD pairs that have macro-paths transversing through reservoirs 1 and 3 are more likely to deviate from the UE principle during peak hours. This can be due to network saturation during peak hours in these two reservoirs, which can be observed from estimated MFDs in Fig. 9. It is also noticed that OD pairs which have fewer number of trips have higher gap values. However, this can be due to a lack of robust and representative data.

At the same time, the objective of the present analysis is not to classify OD pairs that satisfy the UE conditions. Moreover, these gaps can change if a finer spatial partition is used. The reservoirs in the present partition have relatively large areas, where aggregation on macroscopic variables can introduce biases in the gap values. A network partition with smaller reservoirs results in less aggregation and can provide a better idea regarding the UE principle in the network. However, this is out of the scope of the present work. The objective of this study is to show that phone data can be used to extract useful information like the UE gap, which is otherwise only possible with a DTA simulation. This type of data can also be used to validate DTA simulations and calibrate the input parameters in the simulation framework.

9. Conclusions

The present work proposed a framework to estimate the parameters of *all* required components to perform a multi-region MFD model simulations using mobile phone data. The methodology used to select the data and segment individual data records into representative trips was illustrated. Since LBS data was used in the current work, the frequency of data collection varied widely resulting in trips with very sparse records. A method capable of enriching these types of trip using a map matching scheme was discussed in-detail.

A simple partition of the Dallas city network was then considered to estimate the macroscopic variables. It is important to state that the proposed framework is independent of the partitioning scheme and can be used with any other network partition. Firstly, the error of the proposed trip enhancement scheme was estimated using the set of high resolution trips from the *raw data*. It was concluded from the relative errors that the trip enrichment scheme introduced few or acceptable errors in the traveled distances. The next step consisted in estimating the penetration rates of probe vehicles from the data and to this end, OD matrix data from the city council, LDD and the present LBS data were fused together to obtain time dependent penetration rates. Two different types of penetration rates were proposed, namely the OD specific penetration rate and the origin specific penetration rate. As the names suggest, the OD specific rate takes into account the OD flow, while the origin specific rate was computed based on the flow that originated within a zone irrespective of its destination. Using the estimated time dependent penetration rates, the mean density and mean flow of different reservoirs were computed to estimate the MFD for each day separately. An in-depth discussion on temporal and spatial variation of penetration rates was presented. Only days that showed MFDs with similar characteristics were selected and a mean MFD was estimated. It was seen that the MFDs for all the reservoirs were reasonably well-defined with low scatter. It was also observed that the MFDs computed from both the OD specific and the origin specific penetration rates were very similar. It is noteworthy that the penetration rates in the present framework were not constant and were time-dependent. The following part of the work presented the analysis of the trip lengths between the macroscopic OD pairs. This type of analysis is only possible with phone data due to the massive number of records available. The importance of considering the average trip length per macro-path instead of using mean trip length per reservoir was discussed. The evolution of dynamic trip lengths that depends on the traffic conditions was discussed in-detail. This work was first to show the empirical evidence of existence of well-defined relationship between the trip detour ratio (surrogate to mean trip lengths) and the mean speeds at the regional level. Finally, in the third part the evolution of the path flow distributions and UE gaps was discussed and the evolution of the mean path flow coefficient for different OD pairs was illustrated. It could be seen that the path flow coefficient in several OD pairs exhibited a strong variation during the morning and evening peak hours. Similarly, the evolution of the UE gap was estimated based on travel time information. This gap parameter can be used to observe how far the network is from the UE and as well, it can be used to validate the DTA simulations. It was also noticed that the UE gap varied with time during a typical day scenario and this variation was explained for a few OD pairs. Finally, the proportion of OD pairs that satisfied the given threshold UE gap condition was presented. It is noticed that most OD pairs satisfy the UE principle, where relatively low gap values are observed. However, there are few OD pairs that deviate from this principle especially, when OD pairs have macro-paths that transverse through congested reservoirs.

The framework proposed is very generic and network independent. It can be applied to any network that has sufficient phone data and all the parameters necessary to perform a MFD simulation can be calibrated. In the case of the absence of OD matrix data, they can be replaced by the census data of the network to estimate the penetration rates. Overall, the present framework can estimate many interesting and useful parameters that can be used to both perform MFD simulations and validate them. Most of the analysis presented in the current framework cannot be performed using other types of data sources and it is worth emphasizing once again that trip length and path flow distribution analysis can only be achieved through access to massive phone data.

CRediT authorship contribution statement

Mahendra Paipuri: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Yanyan Xu:** Software, Data curation. **Marta C. González:** Writing - review & editing, Formal analysis, Supervision, Project administration, Funding acquisition. **Ludovic Leclercq:** Conceptualization,

Methodology, Formal analysis, Investigation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgment

M. Paipuri and L. Leclercq are grateful for funding from the European Research Council (ERC) in the framework of the European Union's Horizon 2020 research and innovation program (grant agreement No 646592 – MAGnUM project). Y. Xu and M. González would like to thank the Berkeley Deep Drive (BDD) consortium for its support.

Appendix A. Evolution of detour ratio

This section presents an analysis of the trip detour ratio for Dallas. Often straight line Euclidean distances between origins and destinations are used as proxies for traveled distance (Gutiérrez and García-Palomares, 2008; Boscoe et al., 2012). Recently, Yang et al. (2018a) showed that the ratio between actual trip distance and Euclidean straight line distance is constant for different cities of varying scale in China using taxi GPS data. The present analysis aims to verify if the ratio remains the same for Dallas.

Two types of detour ratios are defined in this context, namely the detour ratio based on the Euclidean distance (d_R^E) and the detour ratio based on the network shortest-path distance (d_R^S). The Euclidean detour ratio (d_R^E) is defined as the ratio of trip distance to Euclidean distance (or straight line distance) (d_E), where the network shortest path detour ratio (d_R^S) is defined as the ratio of trip distance to network shortest-path distance (d_S). Fig. 27 presents the variation of two types of detour ratio for the present set of trips. Interestingly, the non-linear decaying relationship proposed in Yang et al. (2018a) between the detour ratio and the Euclidean distance holds true in the present case as well. However, second-degree terms must be included in the functional form to calibrate a better fit in the present case. The fit coefficients for both cases can be expressed as,

$$\begin{aligned} d_R^E &= 1.290 + 0.141\frac{1}{d_E} + 0.104\frac{1}{d_E^2}, \\ d_R^S &= 1.163 + 0.026\frac{1}{d_S} + 0.130\frac{1}{d_S^2}, \end{aligned} \tag{11}$$

where relatively tight fits are obtained with $R^2 = 0.99$. As $d_E \to 0$, the Euclidean detour ratio reaches 1.29, which is in agreement with the values reported in the literature. On the contrary, the asymptotic value of the network shortest path distance detour ratio is 1.163. This is an expected inference as the network shortest distance is a proxy closer to the actual travel distance than the Euclidean distance. However, more data from other cities with different topologies must be studied to verify if this relation is universal.

Appendix B. Day-to-day MFDs

As described earlier, MFDs are calibrated for each day separately and the days that show similar MFDs are selected. Using the selected days for each reservoir, mean MFDs are estimated and are shown in Figs. 9, 9. In this section, day-to-day MFDs are plotted for selected days to analyze the daily behavior of MFD shapes in different reservoirs. As the majority of the flow is observed in reservoirs 1, 2 and 3, and the daily MFDs of only these reservoirs are plotted.

Fig. 28 presents the estimated MFDs for reservoir 1 for different days using the OD-specific penetration rate within the period of time considered. It is clear that the scatter in the MFDs is relatively larger than the mean MFD shown in Fig. 9a. At the same time, it should be noted that the penetration is both time and space-dependent and thus the scatter in the MFD is justified. On the contrary, if a constant mean penetration rate is used to estimate the MFDs, less scatter is expected in the MFD. Nevertheless, the maximum mean flows and densities for all the days are approximately the same, which suggests that a robust MFD is estimated.

Similarly, Figs. 29 and 30 show the daily MFD for reservoirs 2 and 3, respectively, using OD specific penetration rates. As in the case of reservoir 1, slightly larger scatter in the plots can be noticed compared to their respective mean MFDs.

References

Alonso, B., Ibeas, A., Musolino, G., Rindone, C., Vitetta, A., 2019. Effects of traffic control regulation on network macroscopic fundamental diagram: a statistical analysis of real data. Transp. Res Part A: Policy Pract. 126, 136–151. https://doi.org/10.1016/j.tra.2019.05.012.. http://www.sciencedirect.com/science/article/pii/S0965856418303665.

Ambühl, L., Loder, A., Menendez, M., Axhausen, K.W., 2017. Empirical Macroscopic Fundamental Diagrams: New insights from loop detector and floating car data. In: Transportation Research Board 96th Annual Meeting Transportation Research Board, Washington, doi: 10.3929/ethz-b-000118755. https://www.research-collection.ethz.ch/handle/20.500.11850/167171.

Ambühl, L., Loder, A., Zheng, N., Axhausen, K.W., Menendez, M., 2019. Approximative network partitioning for MFDs from stationary sensor data. Transp. Res. Rec. https://doi.org/10.1177/0361198119843264. in press.

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for Macroscopic Fundamental Diagram estimation. Transp. Res. Part C: Emerg. Technol. 71, 184–197. https://doi.org/10.1016/j.trc.2016.07.013.. http://www.sciencedirect.com/science/article/pii/S0968090X16301267.

Ampountolas, K., Kouvelas, A., 2015. Real-time estimation of critical vehicle accumulation for maximum network throughput. In: 2015 American Control Conference (ACC), pp. 2057–2062. doi:10.1109/ACC.2015.7171036.

Ampountolas, K., Zheng, N., Geroliminis, N., 2017. Macroscopic modelling and robust control of bi-modal multi-region urban road networks. Transp. Res. Part B: Methodol. 104, 616–637. https://doi.org/10.1016/j.trb.2017.05.007.. http://www.sciencedirect.com/science/article/pii/S0191261515300370.

Batista, S., Leclercq, L., Geroliminis, N., 2019. Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. Transp. Res. Part B: Methodol. 122, 192–217. https://doi.org/10.1016/j.trb.2019.02.009.. http://www.sciencedirect.com/science/article/pii/S0191261518311603.

Batista, S.F.A., Leclercq, L., 2019. Regional dynamic traffic assignment framework for macroscopic fundamental diagram multi-regions models. Transp. Sci. 53,

1563–1590. https://doi.org/10.1287/trsc.2019.0921.
Batista, S.F.A., Leclercq, L., Menendez, M., 2020. Dynamic Traffic Assignment for regional networks with traffic-dependent trip lengths and regional paths. Transp. Res. Part C: Emerg. Technol Submitted for publication.
Bazzani, A., Giorgini, B., Gallotti, R., Giovannini, L., Marchioni, M., Rambaldi, S., 2011. Towards congestion detection in transportation networks using GPS data. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 1455–1459. https://doi.org/10.1109/PASSAT/SocialCom.2011.249.
Beibei, J.Y., Xu, M., Li, J., van Zuylen, H.J., 2018. Determining the macroscopic fundamental diagram from mixed and partial traffic data. Promet - Traffic Transp. 30, 267–279. https://doi.org/10.7307/ptt.v30i3.2406.. https://traffic.fpz.hr/index.php/PROMTT/article/view/2406.
Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004.. http://www.sciencedirect.com/science/article/pii/S0198971516303970.
Boscoe, F.P., Henry, K.A., Zdeb, M.S., 2012. A nationwide comparison of driving distance versus straight-line distance to hospitals. Profess. Geogr. 64, 188–196. https://doi.org/10.1080/00330124.2011.583586.
Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of Macroscopic Fundamental Diagrams. Transp. Res. Rec.: J. Transp. Res. Board 2124, 127–136, doi:10.3141/2124-12.
Cao, J., Menendez, M., 2015. System dynamics of urban traffic based on its parking-related-states. Transp. Res. Part B: Methodol. 81, 718–736. https://doi.org/10.1016/j.trb.2015.07.018.. iSTTT 21 for the year 2015. http://www.sciencedirect.com/science/article/pii/S0191261515001654.
Chen, J., Bierlaire, M., 2015. Probabilistic multimodal map matching with rich smartphone data. J. Intell. Transp. Syst. 19, 134–148, doi:10.1080/15472450.2013.764796.
Courbon, T., Leclercq, L., 2011. Cross-comparison of Macroscopic Fundamental Diagram estimation methods. Procedia – Soc. Behav. Sci. 20, 417–426. https://doi.org/10.1016/j.sbspro.2011.08.048.. http://www.sciencedirect.com/science/article/pii/S1877042811014285.
Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transp. Res. Part B: Methodol. 41, 49–62. https://doi.org/10.1016/j.trb.2006.03.001.. http://www.sciencedirect.com/science/article/pii/S0191261506000282.
Du, J., Rakha, H., Gayah, V.V., 2016. Deriving macroscopic fundamental diagrams from probe data: issues and proposed solutions. Transp. Res. Part C: Emerg. Technol. 66, 136–149. https://doi.org/10.1016/j.trc.2015.08.015.. advanced Network Traffic Management: From dynamic state estimation to traffic control. http://www.sciencedirect.com/science/article/pii/S0968090X15003162.
Du, J., Rakha, H.A., 2019. Constructing a network fundamental diagram: synthetic origin–destination approach. Transp. Res. Rec. https://doi.org/10.1177/0361198119851445. in press.
Edie, L.C., 1963. Discussion of traffic stream measurements and definitions. In: The 2nd International Symposium on Theory of Traffic flow, London.
Feng, T., Timmermans, H.J., 2016. Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. Transp. Plann. Technol. 39, 180–194. https://doi.org/10.1080/03081060.2015.1127540.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transp. Res. Part B: Methodol. 42, 759–770. https://doi.org/10.1016/j.trb.2008.02.002.. http://www.sciencedirect.com/science/article/pii/S0191261508000180.
Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transp. Res. Part B: Methodol. 45, 605–617. https://doi.org/10.1016/j.trb.2010.11.004.. http://www.sciencedirect.com/science/article/pii/S0191261510001372.
Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A GPS/GIS method for travel mode detection in new york city. Comput. Environ. Urban Syst. 36, 131–139. https://doi.org/10.1016/j.compenvurbsys.2011.05.003.. special Issue: Geoinformatics 2010. http://www.sciencedirect.com/science/article/pii/S0198971511000536.
Gu, Z., Shafiei, S., Liu, Z., Saberi, M., 2018. Optimal distance- and time-dependent area-based pricing with the Network Fundamental Diagram. Transp. Res. Part C: Emerg. Technol. 95, 1–28. https://doi.org/10.1016/j.trc.2018.07.004.. http://www.sciencedirect.com/science/article/pii/S0968090X18300573.
Gutiérrez, J., García-Palomares, J.C., 2008. Distance-measure impacts on the calculation of transport service areas using GIS. Environ. Plann. B: Plann. Des. 35, 480–503. https://doi.org/10.1068/b33043.
Haddad, J., Mirkin, B., 2017. Coordinated distributed adaptive perimeter control for large-scale urban road networks. Transp. Res. Part C: Emerg. Technol. 77, 495–515. https://doi.org/10.1016/j.trc.2016.12.002.. http://www.sciencedirect.com/science/article/pii/S0968090X16302509.
Hagberg, A., Swart, P., S Chult, D., 2008. Exploring network structure, dynamics, and function using NetworkX.
Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp. Res. Part C: Emerg. Technol. 18, 568–583. https://doi.org/10.1016/j.trc.2009.10.006.. http://www.sciencedirect.com/science/article/pii/S0968090X09001430.
Hsieh, H.P., Li, C.T., L, S., 2015. Measuring and recommending time-sensitive routes from location-based data. In: Wooldridge, M., Yang, Q. (Eds.), IJCAI 2015 – Proceedings of the 24th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, pp. 4193–4196.
Huang, C., Zheng, N., Zhang, J., 2019. Investigation of bimodal macroscopic fundamental diagrams in large-scale urban networks: empirical study with GPS data for Shenzhen city. Transp. Res. Rec. 2673, 114–128. https://doi.org/10.1177/0361198119843472.
Ji, Y., Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transp. Res. Part B: Methodol. 46, 1639–1656. https://doi.org/10.1016/j.trb.2012.08.005.. http://www.sciencedirect.com/science/article/pii/S0191261512001099.
Jie, L., van Zuylen, H., Chunhua, L., Shoufeng, L., 2011. Monitoring travel times in an urban network using video, GPS and bluetooth. Procedia – Soc. Behav. Sci. 20, 630–637, doi: 10.1016/j.sbspro.2011.08.070. http://www.sciencedirect.com/science/article/pii/S1877042811014509. the State of the Art in the European Quantitative Oriented Transportation and Logistics Research – 14th Euro Working Group on Transportation & 26th Mini Euro Conference & 1st European Scientific Conference on Air Transport.
Jin, P.J., Cebelak, M., Yang, F., Zhang, J., Walton, C.M., Ran, B., 2014. Location-based social networking data: Exploration into use of doubly constrained gravity model for Origin-Destination estimation. Transp. Res. Rec. 2430, 72–82, doi:10.3141/2430-08.
Kavianipour, M., Saedi, R., Zockaie, A., Saberi, M., 2019. Traffic state estimation in heterogeneous networks with stochastic demand and supply: Mixed Lagrangian-Eulerian approach. Transp. Res. Rec. 2673, 114–126. https://doi.org/10.1177/0361198119850472.
Knoop, V.L., Hoogendoorn, S.P., 2014. Network transmission model: a dynamic traffic model at network level, in: Transportation Research Board 93rd Annual Meeting Transportation Research Board, Washington. http://resolver.tudelft.nl/uuid:8c838988-a7e5-4164-a5b4-ead0160081cf.
Knoop, V.L., van Erp, P.B.C., Leclercq, L., Hoogendoorn, S.P., 2018. Empirical MFDs using google traffic data. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3832–3839. https://doi.org/10.1109/ITSC.2018.8570005.
Kouvelas, A., Saeedmanesh, M., Geroliminis, N., 2017. Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. Transp. Res. Part B: Methodol. 96, 26–45. https://doi.org/10.1016/j.trb.2016.10.011.. http://www.sciencedirect.com/science/article/pii/S019126151630710X.
Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transp. Res. Part B: Methodol. 62, 1–12. https://doi.org/10.1016/j.trb.2014.01.007.. http://www.sciencedirect.com/science/article/pii/S0191261514000174.
Leclercq, L., Paipuri, M., 2019. Macroscopic traffic dynamics under fast-varying demand. Transp. Sci. 53, 1526–1545. https://doi.org/10.1287/trsc.2019.0908.
Leclercq, L., Sénécat, A., Mariotte, G., 2017. Dynamic macroscopic simulation of on-street parking search: a trip-based approach. Transp. Res. Part B: Methodol. 101, 268–282. https://doi.org/10.1016/j.trb.2017.04.004.. http://www.sciencedirect.com/science/article/pii/S0191261516309717.
Mahmassani, H., Williams, J., Herman, R., 1984. Investigation of network-level traffic flow relationships: some simulation results. Transp. Res. Rec. 121–130.
Mariotte, G., Leclercq, L., Batista, S., Krug, J., Paipuri, M., 2020. Calibration and validation of multi-reservoir MFD models: a case study in Lyon. Transp. Res. Part B: Methodol. 136, 62–86. https://doi.org/10.1016/j.trb.2020.03.006.. http://www.sciencedirect.com/science/article/pii/S0191261519306769.
Mohajerpoor, R., Saberi, M., Vu, H.L., Garoni, T.M., Ramezani, M., 2019. H_∞ robust perimeter flow control in urban networks with partial information feedback. Transp. Res. Part B: Methodol. http://www.sciencedirect.com/science/article/pii/S0191261518308609, doi: 10.1016/j.trb.2019.03.010. (In press).
Montoya, D., Abiteboul, S., Senellart, P., 2015. Hup-me: Inferring and reconciling a timeline of user activity from rich smartphone data. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Association for Computing Machinery, New York, NY, USA, doi:10.1145/

2820783.2820852.
Mun, M.Y., Estrin, D., Burke, J., Hansen, M.H., 2008. Parsimonious mobility classification using GSM and WiFi traces. In: Proceedings of the Fifth Workshop on Embedded Networked Sensors.
Nagle, A.S., Gayah, V.V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. Transp. Res. Rec. 2421, 1–11, doi:10.3141/2421-01.
NCTCOG, The North Central Texas Council of Governments. https://www.nctcog.org/. Accessed: 2019-09.
Nikolić, M., Bierlaire, M., 2017. Review of transportation mode detection approaches based on smartphone data. In: 17th Swiss Transport Research Conference, Ascona, Switzerland.
Reddy, S., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2008. Determining transportation mode on mobile phones. In: 2008 12th IEEE International Symposium on Wearable Computers, pp. 25–28.
Saeedmanesh, M., Geroliminis, N., 2016. Clustering of heterogeneous networks with directional flows based on "Snake" similarities. Transp. Res. Part B: Methodol. 91, 250–269. https://doi.org/10.1016/j.trb.2016.05.008.. http://www.sciencedirect.com/science/article/pii/S0191261515302605.
Sbayti, H., Lu, C.C., Mahmassani, H.S., 2007. Efficient implementation of method of successive averages in simulation-based dynamic traffic assignment models for large-scale network applications. Transp. Res. Rec. 2029, 22–30, doi:10.3141/2029-03.
Shim, J., Yeo, J., Lee, S., Hamdar, S.H., Jang, K., 2019. Empirical evaluation of influential factors on bifurcation in macroscopic fundamental diagrams. Transp. Res. Part C: Emerg. Technol. 102, 509–520. https://doi.org/10.1016/j.trc.2019.03.005.. http://www.sciencedirect.com/science/article/pii/S0968090X18304042.
Shoufeng, L., Jie, W., van Zuylen, H., Ximin, L., 2013. Deriving the macroscopic fundamental diagram for an urban area using counted flows and taxi GPS. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 184–188. doi:10.1109/ITSC.2013.6728231.
Sinnott, R.W., 1984. Virtues of the haversine. Sky Telescope 68, 158–159.
Sirmatel, I.I., Geroliminis, N., 2018. Economic model predictive control of large-scale urban road networks via perimeter control and regional route guidance. IEEE Trans. Intell. Transp. Syst. 19, 1112–1121.
Tsubota, T., Bhaskar, A., Chung, E., 2014. Macroscopic Fundamental Diagram for Brisbane, Australia: Empirical findings on network partitioning and incident detection. Transp. Res. Rec. 2421, 12–21, doi:10.3141/2421-02.
Wang, P., Wada, K., Akamatsu, T., Hara, Y., 2015. An empirical analysis of macroscopic fundamental diagrams for sendai road networks. Interdiscip. Inf. Sci. 21, 49–61. https://doi.org/10.4036/iis.2015.49.
Xu, Y., Clemente, R.D., González, M.C., 2020. Understanding route choice behavior with location-based services data. Transp. Res. Part C: Emerg. Technol Submitted for publication.
Yang, H., Ke, J., Ye, J., 2018a. A universal distribution law of network detour ratios. Transp. Res. Part C: Emerg. Technol. 96, 22–37. https://doi.org/10.1016/j.trc.2018.09.012.. http://www.sciencedirect.com/science/article/pii/S0968090X18311185.
Yang, X., Stewart, K., Tang, L., Xie, Z., Li, Q., 2018b. A review of GPS trajectories classification based on transportation mode. Sensors (Basel, Switzerland) 18, 3741. https://doi.org/10.3390/s18113741.. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6263992/.
Yildirimoglu, M., Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. Transp. Res. Part B: Methodol. 70, 186–200. https://doi.org/10.1016/j.trb.2014.09.002.. http://www.sciencedirect.com/science/article/pii/S0191261514001568.
Yildirimoglu, M., Ramezani, M., Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. Transp. Res. Part C: Emerg. Technol. 59, 404–420, doi: 10.1016/j.trc.2015.05.009. http://www.sciencedirect.com/science/article/pii/S0968090X15001813. special Issue on International Symposium on Transportation and Traffic Theory.
Yildirimoglu, M., Sirmatel, I.I., Geroliminis, N., 2018. Hierarchical control of heterogeneous large-scale urban road networks via path assignment and regional route guidance. Transp. Res. Part B: Methodol. 118, 106–123. https://doi.org/10.1016/j.trb.2018.10.007.. http://www.sciencedirect.com/science/article/pii/S0191261518301152.
Zhang, L., Dalyot, S., Eggert, D., Sester, M., 2012. Multi-stage approach to travel-mode segmentation and classification of GPS traces. In: ISPRS Guilin 2011 Workshop on International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 87–93.