Impact of data processing on deriving micro-mobility patterns from vehicle availability data

Journal Article

Author(s):
Zhao, Pengxiang; Haitao, He; Li, Aoyong; Mansourian, Ali

Publication date:
2021-08

Permanent link:
https://doi.org/10.3929/ethz-b-000491982

Rights / license:
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

Originally published in:
Impact of data processing on deriving micro-mobility patterns from vehicle availability data

Pengxiang Zhao\textsuperscript{a}, He Haitao\textsuperscript{b, *}, Aoyong Li\textsuperscript{c, *}, Ali Mansourian\textsuperscript{a, d}

\textsuperscript{a} GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
\textsuperscript{b} School of Architecture, Building and Civil Engineering, Loughborough University, UK
\textsuperscript{c} Institute for Transport Planning and Systems (IVT), ETH Zürich, Zürich CH-8093, Switzerland
\textsuperscript{d} Center for Middle-Eastern Studies, Lund University, Lund, Sweden

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Micro-mobility
E-scooter sharing
Data processing
Data sampling
Spatio-temporal patterns
Vehicle availability data
GPS
Trip identification

\textbf{ABSTRACT}

Vehicle availability data is emerging as a potential data source for micro-mobility research and applications. However, there is not yet research that systematically evaluates or validates the processing of this emerging mobility data. To fill this gap, we propose a generally applicable data processing framework and validate its related algorithms. The framework exploits micro-mobility vehicle availability data to identify individual trips and derive aggregate patterns by evaluating a range of temporal, spatial, and statistical mobility descriptors. The impact of data processing is systematically and rigorously investigated by applying the proposed framework with a case study dataset from Zurich, Switzerland. Our results demonstrate that the sampling rate used when collecting vehicle availability data has a significant and intricate impact on the derived micro-mobility patterns. This research calls for more attention to investigate various issues with emerging mobility data processing to ensure its validity for transportation research and practices.

1. Introduction

With the technological development of information and communications technology (ICT) and the emergence of innovative urban mobility services in the past few years, shared micro-mobility has been one of the fastest-growing branches of transportation. As one of the environmentally friendly transport modes, shared micro-mobility services, including dockless e-scooters, dockless and docked bikes/e-bikes, have shown a tantalising potential for improving short- and medium-distance travel (Barbour et al., 2019; Luo et al., 2020; Li et al., 2021; Reck et al., 2021). However, up to now, their usage is still far from reaching equilibrium and is largely driven by unpredictable factors such as the volatile investment environment, disruptive business models, changing policies, and unprecedented events like COVID-19. Therefore, regular monitoring and better understanding of their fast-evolving spatio-temporal patterns are crucial for the effective deployment of the fleets in cities worldwide, both for the city authorities and the service providers. For example, over twenty shared e-scooter trials have launched in the UK since September 2020 and the Department for Transport aims to implement central monitoring and evaluation across all trial areas (GOV.UK, 2020). This is only feasible by exploiting available data.

A variety of traditional data sources have been exploited for such purposes, for example questionnaire surveys (Guo et al., 2017; Ma et al., 2020) and GPS trajectories from micro-mobility providers (Christian et al., 2019; Ji et al., 2020; Li et al., 2020a). These data are costly and difficult to obtain. Some recent studies (Xu et al., 2019; Zhu et al., 2020) have exploited vehicle availability data, namely the

\* Corresponding authors.
E-mail addresses: H.He@lboro.ac.uk (H. Haitao), aoyong.li@ivt.baug.ethz.ch (A. Li).

https://doi.org/10.1016/j.trd.2021.102913
GPS locations of all available vehicles from service providers. This kind of data is generally available from all operators at a city-scale, as it is the basic information provided to the potential users to locate the available vehicles. It can be accessed and collected from operator database, open data repository of some city authorities, and recently also directly from application programming interface (API) of some operators in many European and US cities. Despite existing literature (Wang and Chen, 2018; Hassanpour et al., 2020; Xu et al., 2020) pointing out data quality as a key challenge when using emerging big data for transportation research, surprisingly there is not yet research that systematically evaluates or validates the processing of micro-mobility vehicle availability data. As valid data builds the foundation for valid research, the data processing itself is an important subject to scrutinise.

Unlike GPS trajectories from micro-mobility providers recording the whole process of vehicle movement, vehicle availability data only records the vehicle locations when they are available. Considering the short-distance artificial movement and the oscillation issue caused by GPS drifting, the vehicle availability data must be processed precisely (e.g. identify valid trips) first before it can be used to derive aggregate patterns (e.g. evaluate temporal, spatial, and statistical mobility descriptors). Determining an appropriate sampling rate is another tricky issue in the vehicle availability data collection, which involves a trade-off between data completeness and data redundency. On one hand, it is necessary to collect sufficient data with a high sampling rate to guarantee the completeness and representativeness of the data. On the other hand, if the chosen sampling rate is too high, it will result in data redundency, thereby increasing the data computation and storage costs. Therefore, it is essential to evaluate a sampling rate that is cost-effective yet generating adequately accurate results for the micro-mobility travel patterns. Several studies explored the impact of sampling rate on mobility descriptors with some other data sources (e.g. call detail record) (Ranjan et al., 2012; Zhao et al., 2019; Burkhard et al., 2020). To the authors’ best knowledge, there is not yet research that examines the influence of data sampling rate on micro-mobility trip patterns using vehicle availability data.

To fill the research gap, this research proposes a generally applicable data processing framework and validates its related algorithms. The framework exploits micro-mobility vehicle availability data to identify individual trips and derive aggregate patterns. The impact of data processing is systematically and rigorously investigated by applying the proposed framework with a case study dataset from Zurich, Switzerland. The results demonstrate how vehicle availability data can be used by researchers and practitioners to understand micro-mobility and how data processing influences the results. Even though an e-scooter sharing dataset is tested here as a case study, the framework and recommendations are generally applicable to bike-sharing and other micro-mobility services as well.

The remainder of this paper is organised as follows. Section 2 presents a comprehensive review of shared micro-mobility studies with an emphasis on various data sources. Section 3 describes the data used in this research. Section 4 elaborates the proposed methodology and validate its related algorithms. Section 5 presents the case study results from a systematic and comprehensive experiment. Section 6 summarises the contributions of this work and suggests future research directions.

2. Literature review

Many studies have analysed micro-mobility travel patterns. They can be categorised into three groups according to the data sources, namely survey-based, GPS-based and API-based studies.

The studies based on survey data are mainly concentrated on exploring the influencing factors of bike-sharing from different aspects, such as socio-demographic, urban environment, user preference (Bachand-Marleau et al., 2012; Heesch et al., 2012; Buck et al., 2013; Campbell et al., 2016; Du and Cheng, 2018). Some studies examined the potential substitution of bike-sharing on motor trips (Barbour et al., 2019; Jia and Fu, 2019; Ma et al., 2020). With the fast growing demand and prevalence of e-scooter and e-moped sharing, the number of studies on them via survey data is also increasing (Aguilera-Garcia et al., 2020; Baek et al., 2021). Due to the limited data sample size, the studies on spatio-temporal patterns of micro-mobility trips at the population level are scarce with survey data.

In recent years, there is also an emergence of studies using GPS-based micro-mobility data. For some studies, the researchers directly obtained the origin–destination data or trip data (i.e. including the GPS tracking points between origin and destination) from bike-sharing operators (Li et al., 2020b). For example, Du et al. (2019) developed a framework to explore the spatio-temporal usage patterns of dockless shared bikes using the bike usage data from Mobike. The results uncovered that the trip time and distance statistically follow log-normal distributions, rather than power-law distribution. The study by Li et al. (2020c) displayed a data-driven framework to understand intra-urban human mobility via exploratory spatial, temporal, and statistical analysis of users’ trip data from Mobike. The results revealed that bike-sharing usage showed strong diurnal patterns (i.e. two peaks on weekdays and one peak on weekends) in time, while its spatial distribution varied remarkably at different times of a day. Ji et al. (2020) performed a comparison study on usage regularity and its determinants between docked and dockless bike-sharing systems using smart card data of a docked bike-sharing scheme and GPS trajectory data of a dockless bike-sharing scheme in Nanjing, China. It was found that the trips during morning and afternoon peak hours were positively related to the usage regularity of two bike-sharing systems. Xing et al. (2020) explored the travel patterns and trip purposes of dockless bike-sharing using massive trip data, which was collected from the Mobike company in Shanghai, China. However, these trajectory data are still difficult to obtain from operators.

The third group of studies used data collected from mobile applications or websites of micro-mobility systems via a web crawler or an open API. This data collection method has previously been used to study free-floating car sharing (Hughes and Mackenzie, 2016; Wang et al., 2017; Cooper et al., 2018; Hassanpour et al., 2020). With the boom of micro-mobility services, this type of data is also used in micro-mobility studies. For instance, Faghfih-Imani et al. (2017) conducted an empirical analysis of bike-sharing usage and rebalancing by capturing bike-sharing system state data from Barcelona and Seville, Spain via the websites of these bike-sharing programs every 5 min. The results indicated that bicycle infrastructure attributes and land-use characteristics are crucial to bicycle usage and operator rebalancing. Reynaud et al. (2018) developed a behaviourally quantitative model to examine bike availability at a station
using Montreal BIXI data, which were collected from BIXI Montreal’s website on a minute basis. Xu et al. (2019) uncovered the patterns of cycling activities using a bike-sharing dataset over four months in Singapore, which were collected from a dockless bike-sharing operator with a frequency of 5 min on average. It was reported that the temporal usage patterns of bikes displayed structural variations across urban locations. Yang et al. (2019) investigated the changes in travel behaviours by capturing available bikes locations every four minutes and analysing bike-sharing during a period when a new metro line came into operation in Nanchang, China. The results showed how the spatio-temporal patterns of bike travel behaviour changed over the period. McKenzie (2020) conducted a comparison study to explore the spatial and temporal similarities and differences of usage patterns between the mobility companies by accessing the data every minute via open API.

Although vehicle availability data is increasingly available from open APIs of micro-mobility operators, there is not yet research that systematically evaluates or validates the impact of data processing on modeling micro-mobility trip patterns. Hence, it is important to bridge this gap.

3. Vehicle availability data collection

We collect data in Zurich, Switzerland. Several micro-mobility operators provide services in the study area, including shared bikes, e-bikes, and e-scooters. Vehicle availability data from one dockless e-bike operator was collected on 1st and 2nd of February, 2020 to validate the proposed methods, because the booking data of actual trips is available from this operator as the ground truth. Vehicle availability data from dockless e-scooter operators was collected from February 1st to February 29th 2020 to examine the impact of data processing on deriving mobility patterns. Both datasets are obtained by scanning the available vehicles in each area every 30 s on average. As displayed in Table 1, each record contains the fields including id, vehicle id, timestamp, longitude, latitude, and provider. The e-scooter data over one month consists of about 148 million such records.

Note that none of the operators provide reservation services at the moment. Users need to unlock the vehicle before starting the trip. Once a vehicle is unlocked, it is no longer available in the system and cannot be scanned. Since the periods when vehicles are rented out have not been recorded, the collected data can only reflect the origin and destination of each trip.

4. Methodology

We develop the following framework to examine the effect of sampling rate on modeling micro-mobility trip patterns from vehicle availability data. As shown in Fig. 1, the framework comprises four main steps, namely data sampling, trip identification, evaluation of micro-mobility descriptors, and measurement of the effect.

4.1. Data sampling

To examine the effect of sampling rate on typical human mobility indicators, Zhao et al. (2019) defined the temporal sampling interval (TSI) as the interval between pairs of consecutive records in mobile phone data. Likewise, this study introduces TSI to measure and depict the temporal sampling rate of micro-mobility vehicle availability data. To quantify the effects of sampling rate on micro-mobility descriptors, we sample vehicle availability data with different temporal intervals. The dataset for each interval will be further utilised to calculate the descriptors. The original sampling rate of raw data (i.e. 30 s on average), could be regarded as a benchmark. Then, a downsampling method is implemented to resample the dataset with different TSIs (e.g. from 2 min to n minutes). We denote the benchmark dataset as $D_1$, the other downsampled datasets are expressed as $D_2, \ldots, D_n$ successively according to the selected TSI.

Specifically, given a start time $t_0$ and data sampling interval $tsi$, the record $r$ with timestamp that is closest to $t_0 + ts$ and higher than $t_0 + ts$ will be chosen. Then the timestamp of record $r$ will be set as the start time to sample the next record in the same way until $t_0 + tsi$ is higher than the timestamp of the last point in the data. Ultimately, $n$ groups of datasets are generated by downsampling.

Due to the influence of some external factors (e.g. signal loss), the data missing is inevitable while collecting the data. Hence, even if the sampling rate is designated before data collection, there are still intervals between two adjacent records that are more than the determined sampling interval. To guarantee a high sampling frequency, we filter the e-scooters and keep the data of those with the recording time during one day exceeding 12 h and the total number of records being more than 720. It implies that the data is required to be recorded at least half day and the average sampling interval is no less than one minute. Taking the benchmark dataset on 1st February as an example, most of the intervals are less than two minutes, which accounts for 99.2% of the intervals. More than 75% of intervals are within one minute, but no sampling intervals equal to zero. It should be noted that the sampling is not stable although the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Samples of vehicle availability data records.</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>vehicle id</td>
</tr>
<tr>
<td>756384199</td>
<td>35358</td>
</tr>
<tr>
<td>756384201</td>
<td>37715</td>
</tr>
<tr>
<td>756400672</td>
<td>226704</td>
</tr>
<tr>
<td>756400673</td>
<td>227541</td>
</tr>
<tr>
<td>756638609</td>
<td>7ced57d1-d1fc</td>
</tr>
<tr>
<td>756638610</td>
<td>c7599409-8bac</td>
</tr>
</tbody>
</table>
sampling interval is set as one minute. The actual sampling intervals vary within 60 s, including some sampling intervals less than five seconds. These statistics indicate that the data records are collected intensively.

Previous studies have demonstrated that the average duration of e-scooter trips is around 10 min (Jiao and Bai, 2020; McKenzie, 2020). Hence, 10 min is determined as the maximum sampling interval as it is meaningless to set the sampling interval longer than the duration of trips. Considering that if the resampling size is too small (e.g. 30 s) there will be so many resampled datasets generated, the sampling size is set as one minute in this study. Eventually, 9 groups of resampled datasets are obtained (i.e. $D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}$), which will be compared with the benchmark dataset $D_1$. According to the data sampling description, we down-sample the data during the whole month based on the benchmark dataset $D_1$ by setting the sampling interval from 2 min to 10 min. Fig. 2a presents the number of records for the benchmark dataset and each down-sampled dataset. It is found that a dramatic decrease occurs between $D_1$ and $D_2$ in terms of the number of records. This is because that the sampling frequency is more than twice on average per minute for the benchmark dataset $D_1$. Moreover, we further calculate the record reduction ratio with the following equation:

$$\text{Ratio} = \frac{D_1 - D_i}{D_1}$$ (1)

where $D_1$ and $D_i$ represent the number of records for the benchmark dataset $D_1$ and down-sampled dataset $D_i$ respectively.
Fig. 2b displays the changes of record reduction ratio with the increase of the data sampling interval, which belong to the range of [79%, 95.5%] indicating that a large amount of records are filtered after the data sampling.

4.2. Identifying trips

Identifying trips from raw data is a key step in micro-mobility studies. The reliability of experimental results relies heavily on the accuracy of the identified trips. Since the GPS coordinates \((P_i)\) are recorded only when the vehicle is available (i.e. not rented), the trajectory for each vehicle can be denoted as a sequence of records \(S_1, S_2, \ldots, S_n\) chronologically. The trips can be obtained automatically by identifying the two consecutive stops. However, due to GPS signal drifting, even if one vehicle is parked at a certain location, the recorded coordinates of longitude and latitude over a period can be slightly different. In other words, although a vehicle is not moved, the recorded GPS locations are not the same for the same point. It implies that if we directly identify the trips according to whether the locations of two consecutive records change, some identified short trips might be fake.

To identify the real trips, the GPS records are sorted first according to their collected time first. Then, the geographic distance between every two consecutive records \((P_i, P_{i+1})\) are calculated. If the distance between \(P_i\) and \(P_{i+1}\) is shorter than parameter \(\delta\), the two consecutive records are regarded as the same stop. \(\delta\) is a distance threshold that depends on the positioning error of GPS device. Accordingly, given one vehicle, its records can be divided into multiple classes, each of which corresponds to one stop. The centroid of the records in the same class is calculated as the coordinates of the corresponding stop to mitigate the bias caused by the positioning error of GPS device. After the processing, the sequence of records for one vehicle is represented as a sequence of stops. Each stop corresponds to one origin or destination, which can be transformed into trips ultimately. The whole process of trip identification is described in Algorithm 1 of Appendix A.

As mentioned in Section 3, vehicle availability data from one dockless e-bike operator was collected on 1st and 2nd of February, 2020 to validate the proposed methods, because the booking data of actual trips is available from this operator as the ground truth. By removing the abnormal trips based on the standards, 299 actual trips from 138 e-bikes are obtained eventually from the booking data. Then we calculate the trips from the corresponding vehicle availability data based on the trip identification method. By conducting the positioning error analysis on bike and e-scooter data on 1st of February, as shown in Fig. 3, the positioning error is less than 30 m for both bike-sharing and e-scooter sharing data. Hence, \(\delta\) is set as 30 to identify bike-sharing trips.

After identifying the trips, the following standards are adopted to filter the invalid bike trips based on prior knowledge and cycling speed limit in Switzerland ¹: (1) trip duration is longer than two minutes and less than 2 h, (2) trip distance is greater than 200 m and shorter than 15 km, (3) average speed is less than 25 km/h. Eventually, 304 valid trips are identified from the raw data. To validate the identified trips, the calculated trips are compared with the actual trips in terms of bike id, start and end time as well as pick-up and drop-off points. Thus, one trip \(T\) can be represented as an itemset \(I_T\):

\[
I_T = \{Bid, St, Et, Slon, Slat, Elon, Elat\}
\]

where \(Bid\) is bike id, \(St\) and \(Et\) represent the start and end time, \(Slon\) and \(Slat\) are the longitude and latitude of pick-up point \(o\), \(Elon\) and \(Elat\) are the longitude and latitude of pick-up point \(d\).

Due to the influence of sampling interval and positioning error, there is little possibility that the time and coordinates of the extracted trip and the corresponding actual trips are completely identical. Hence, the thresholds \(\Delta_t\) and \(\Delta_d\) are set for the time and coordinates to check whether two trips are the same. Given one extracted trip \(T_E\) and its corresponding actual trip \(T_A\), the following

¹ https://healthyandsafe.biz/e-scooters/
In this study, the thresholds $\Delta_t$ and $\Delta_d$ are determined as 2 min and 30 meters respectively. After the validation, 294 trips that occupy 98.3% of the actual trips are correctly identified, which demonstrates the reliability of the trip identification method. We further compare the identified and actual trips by taking the number of trips as an example, as shown in Fig. 4. It can be observed that two groups of trips display almost the same temporal patterns. Hence, the developed method of identifying trips from vehicle availability data is completely eligible for this study.

4.3. Mobility modeling

Micro-mobility trip patterns are evaluated from temporal, spatial and statistical perspectives.

4.3.1. Temporal distribution

Whether the temporal usage patterns are regular on weekdays and weekends is one of the main concerns for micro-mobility providers and city planners. We examine the temporal distribution of travel demand on an hourly basis on both weekdays and weekends using the identified trips. To reduce the randomness of the data over one week, the hourly number of trips is averaged by day of the week for the data during one month. Given dataset $D_1$, a temporal signature (TS) is calculated from the average number of trips $N_i$ during each hour, denoting as:

$$\text{TS}_1 = [N_1, N_2, \ldots, N_i, \ldots, N_{24}]$$

(4)

Similarly, the temporal signatures for the downsampled datasets can be constructed accordingly.

4.3.2. Spatial distribution

Spatial distribution of popular destinations of users using micro-mobility services is also beneficial for micro-mobility providers to place and rebalance their vehicles in cities. The drop-off points of trips represent the destinations of citizens who use micro-mobility services for their travel. Based on the identified trips, we evaluate the spatial distribution of destinations using kernel density estimation (KDE). KDE has been widely used to characterise the spatial distribution of geographic points by visualising the data in a smooth density surface, which displays the geographic clustering characteristics of points in two dimensional space (Bailey and Gatrell, 1995; Borruso, 2008). In mathematical terms, KDE is denoted as:

$$f(x, y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{d_i}{h}\right)$$

(5)

Where $f(x, y)$ is the density value at location $(x, y)$. $n$ is the number of points, $h$ represents the bandwidth, $d_i$ is the geographic distance...
between point $i$ and location $(x,y)$. $K$ is a density function.

4.3.3. Statistical distribution

In recent years, statistical methods have been widely used in human mobility studies by quantitatively describing human behaviour events, thereby uncovering the hidden statistical laws and exploring the mechanism of these laws. A plethora of studies demonstrated the scaling laws of human mobility with various human movement data, which are capable of uncovering statistical patterns of human mobility by discovering probability distributions of mobility descriptors (Gonzalez et al., 2008; Song et al., 2010). The three typical mobility descriptors are adopted to examine the statistical characteristics of micro-mobility, which have been widely used in human mobility studies, including trip duration, trip distance, and average speed.

Several commonly used distributions have been used to examine the statistical characteristics of human mobility, such as power law, power law with an exponential cutoff, exponential and Weibull (Liang et al., 2012; Wang et al., 2015; Alessandretti et al., 2017; Barbosa et al., 2018; Zhao et al., 2020). In this study, we apply the three models to our datasets to fit the statistical distribution of the three micro-mobility indicators, including power law, exponential law and Weibull distribution.

4.4. Quantifying the effect of sampling rate on micro-mobility modeling

Based on the downsampled datasets, the mobility descriptors are calculated and compared with those from the benchmark dataset. For the temporal and spatial distribution, similarity measurement methods are used to evaluate the effects of data sampling on the results, including correlation coefficient, and cosine similarity (Tan et al., 2016). For the result of statistical distribution, it is essentially a probability density distribution or cumulative density distribution of a micro-mobility descriptor. Hence, we quantify the effects of data sampling on statistical distribution in two ways. On the one hand, the changes of the distribution exponents are evaluated to measure the effect. On the other hand, Hellinger distance (HD) and Bhattacharyya distance (BD) between the probability distributions are calculated to quantify the effect. In probability and statistics, Hellinger distance and Bhattacharyya distance have been commonly adopted to quantify the similarity between two probability distributions.

5. Results

5.1. The effect of sampling rate on trip identification

The e-scooter trips over the whole month are first identified with the trip identification method. After identifying the trips, data cleaning is implemented to remove the abnormal trips due to some other factors (e.g. artificially rebalance the locations of e-scooters by users or operators). The following standards are adopted to filter the invalid trips based on prior knowledge and e-scooter speed limit in Switzerland: (1) trip duration is more than one minute and less than 2 h, (2) trip distance is greater than 100 m and shorter than 10 km, (3) average speed is less than 20 km/h. The trips that satisfy the above-mentioned requirements will be reserved for the remaining studies.

![Fig. 4. Comparison between the identified and actual trips in terms of the number of trips.](https://healthyandsafe.biz/e-scooters/)
Fig. 5 displays the number of trips before and after the filtering. The number of trips before filtering decreases with the increases of the data sampling interval, which can be attributed to that more short trips would get lost with a longer sampling interval. However, it can be observed that the numbers of trips decrease dramatically after filtering the fake and invalid trips based on the criteria. Through careful scrutiny of the trip count reduction, it is found that the origin and destination of a trip normally change after downsampling the dataset.

Next, the potential explanation is given via a sketch map of how trip changes after the downsampling of vehicle availability data. As shown in Fig. 6a, two red circles represent the real origin and destination of the trip, which starts at \( t_1 \) and ends at \( t_3 \). Fig. 6b displays the corresponding trip after downsampling the dataset. Compared with the trip in the benchmark dataset (Fig. 6a), the origin and destination of the trip in the downsampled dataset might change in time and space (Fig. 6b). In particular, trip duration will become higher while downsampling the dataset. The degree of change for trip distance depends on the location uncertainty due to the position error or artificial shift. Due to the change of trip duration, a number of fake trips (i.e. duration less than one minute) will become qualified while filtering the invalid trips based on the criteria. It explains why the number of trips from \( D_1 \) is smaller than those from other datasets after filtering.

### 5.2. The effect of sampling rate on micro-mobility description

#### 5.2.1. Temporal distribution analysis

To analyse how the temporal distribution of micro-mobility is influenced by the sampling rate, we first examine the hourly variations of the average number of trips using the datasets with different temporal sampling intervals, as shown in Fig. 7. It can be observed that the variations on the average number of trips display different patterns on workday and weekend. The plot by the time of day shows that the usage of e-scooters displays three obvious peaks during morning (i.e. 8:00–9:00) and evening (i.e. 17:00–18:00), which corresponds to the two commuting peaks. It is indicated that the e-scooter usage trips occurring during commuting time occupy a high proportion of the whole trips on workdays. It should be noteworthy that the evening peak is higher than the morning peak indicating more usage of e-scooter after work. Another interesting point of focus is the secondary peak at noon (i.e. 12:00–13:00), which may reflect people’s usage on e-scooter for lunch. While on weekend, the morning and evening peaks as well as lunch peak are disappeared, the peak is shifted to the afternoon (i.e. 16:00–17:00). It also reflects variations on people’s travel demand by e-scooter on weekend that are distinct from that on workday.

As shown in Fig. 7, the visual comparison between the datasets with different sampling rates indicates that the nine groups of resampled datasets produce almost identical temporal distributions with the benchmark dataset \( D_1 \), while the average numbers of trips are slightly overestimated. Moreover, we quantitatively analyse the effect of sampling rate by calculating the correlations between the temporal signature of the benchmark and those of resampled datasets respectively on workday and weekend, as shown in Fig. 8. It is found that all the correlation coefficients reach 0.99 on both workday and weekend. We can conclude that the sampling rate of dataset has a slight influence on the temporal distribution of the number of trips. Even though the influence is tiny, however, the \( R^2 \) values display a decreasing tendency with the increment of the temporal sampling interval.

#### 5.2.2. Spatial distribution analysis

In this section, we further examine how the sampling rate influences the spatial distribution of e-scooter trips. Fig. 9 shows the spatial distributions on the passengers’ trip destinations by e-scooter. The travel density by e-scooter in sub-figures is estimated using kernel density with the kernel function as Gaussian and the bandwidth (or search radius) as 100 m. The red colour represents high density while the blue colour represents low density. To avoid the heat maps occupying too much space, we only take the KDE maps calculated from the datasets \( D_1, D_5 \), and \( D_{10} \) as examples.

By comparing the high density areas of e-scooter usage with the spatial distribution of point of interest (POI), we can conclude that the drop-off points are mainly concentrated on transit stations and some commercial areas. For examples, in Fig. 9, the white circles represent some railway stations, including Bahnhof Oerlikon, Bahnhof Hardbrucke, Bahnhof Altstetten, and Zurich Hauptbahnhof. These railway stations, as the transportation hubs of Zurich, carry enormous daily passenger flow. The black circles in Fig. 9 represent the commercial areas.

For each data, the density of drop-off points around the railway stations is higher than that in the commercial areas on workdays. On weekends, the density of commercial area becomes higher, such as the two commercial areas in the sub-figures. It denotes that the trip purposes on workdays are different from that on weekends. Furthermore, there are no significant differences between the heat maps of \( D_5 \) and \( D_{10} \) with the benchmark \( D_1 \). And thus, we can deduce that the decreased sampling frequency will not lead to various conclusions in terms of the spatial distribution of e-scooter trips.

Next, we quantify the effects of sampling rate on the spatial distribution of travel demand by e-scooter. Specifically, the minimum bounding rectangle of the whole study area is divided into \( 0.005' \times 0.005' \) grids, the trips on workdays and weekends over one month are aggregated to the corresponding grids based on the drop-off points respectively. Then the number of drop-off points is calculated for each grid, and a matrix is generated eventually. As conducted in the temporal distribution section, we also measure the similarity of the matrices that are calculated using the datasets (i.e. from \( D_1 \) to \( D_{10} \)). By calculating the cosine similarity (Singhal et al., 2001) between the benchmark dataset \( D_1 \) and each resampled dataset, the similarity results are displayed in Table 2. The second and third columns in Table 2 indicate that the spatial matrices from different datasets present high correlation with that of the benchmark dataset. Note that the slightly decreasing tendency can still be observed with the increment of the sampling interval, even though all the correlation coefficients yield above 0.97. In general, the low sampling rate has a tiny influence on understanding the spatial
distribution of e-scooter trips when the temporal sampling interval is no more than 10 min.

5.2.3. Statistical distribution analysis

In this section, the effect of sampling rate on the statistical distribution of micro-mobility descriptors is investigated. First, the basic statistics of the selected three micro-mobility descriptors, namely trip duration, distance and average speed, are calculated and displayed with boxplot, as shown in Fig. 10. It can be observed that the sampling rate has different influences on the three indicators. With regards to trip duration, the maximum normal values are around 15 min for all the datasets. However, the other four statistics, including the minimum, the median, and the first and third quartiles, present an increasing tendency overall with the increment of the temporal sampling intervals. The mean and median of trip duration from dataset \(D_1\) are 7.4 and 6.0 min respectively. It is noteworthy that the changes become remarkable when the temporal sampling interval is more than 5 min. Regarding trip distance, it is uncovered that the sampling intervals have no obvious influence on the five statistics of boxplot. The mean and median of trip distance from dataset \(D_1\) yield 0.84 and 0.66 km respectively. In terms of average speed, it is illustrated that the minimum normal values are almost identical for all the datasets. The other four statistics, including the maximum, the median, and the first and third quartiles, tend to
decrease while the temporal sampling intervals are increasing. Second, the statistical characteristics of trip distance, duration and average speed are evaluated by plotting their probability density distribution and cumulative density distribution. As shown in Fig. 11, the probability density distribution and cumulative density distribution of three micro-mobility descriptors are illustrated to evaluate the scaling properties. The benchmark dataset $D_1$ and five groups of the resampled datasets $D_2$, $D_4$, $D_6$, $D_8$ and $D_{10}$ are selected to obtain and visualise the statistical characteristics of e-scooter trips. The first and second columns in Fig. 11 display the probability distribution and cumulative probability distribution of trip duration and distance. It is evident that the trip duration and distance in e-scooter sharing follow right-skewed or heavy-tailed distribution, which is consistent with the phenomenon in bike-sharing (Du et al., 2019). It might be attributed to the same role of e-scooter sharing with bike-sharing in public transportation, which are both mainly used for short-distance trips. Based on the probability distribution of trip duration, it can be observed that the highest proportion of trip duration moves towards right with the increase of sampling interval. This is self-evident that the trips with short duration are disappeared while decreasing the sampling rates. Interestingly, the probability distribution of trip distance displays that the proportion of trip distance between 100 m and 200 m increases after downsampling the dataset. Additionally, it should be noted that, with the increase of data sampling interval, the cumulative probability distribution curves of trip duration changes increasingly remarkably at the start point, while the curves of trip distance change slightly. Regarding the statistical distribution of average speed, it follows the normal distribution approximately in datasets $D_1$ and $D_2$. With the increment of the temporal sampling interval, the normal distribution is gradually transformed to right-skewed distribution, as shown in the third column of Fig. 11.

To further quantitatively compare the changes of statistical distribution with the increment of sampling interval, we fit the distributions with the introduced models in Section 4.3.3. By estimating the statistical distribution of the three micro-mobility indicators, the trip duration, distance and average speed can be fitted by power law, exponential function, and Weibull distribution with high goodness respectively. The estimated parameters are displayed in Table 3.

Furthermore, we quantify the sampling rate effects by measuring the similarity of probability distributions between the benchmark dataset $D_1$ and each resampled dataset. Specifically, the two types of distance introduced in Section 4.4, namely Hellinger distance and Bhattacharyya distance, are calculated to measure the similarity of probability distributions, as shown in Table 4. It can be seen that the Hellinger distance and Bhattacharyya distance between the benchmark dataset and the downsampled data both gradually increase with the increment of sampling interval for duration and speed. With regards to trip distance, the Hellinger distance and Bhattacharyya distance fail to display an increasing tendency while lowering the sampling rate. Hence, we can conclude that sampling rate presents the influence on the statistical distribution of micro-mobility distribution.

5.3. Comparison on computation efficiency

In this section, we compare the computation efficiency in terms of identifying trips from different datasets to further quantify the effect of sampling rate on micro-mobility. Identifying trips from raw data occupies a high proportion of computation time in the whole experiment. Here, the experiment is conducted by taking the data on February 1st as an example, as shown in Table 5. The second and third columns display the number of records and the calculation time of identifying trips for each group of dataset. It is found that the number of records decreases dramatically with the increment of sampling interval, and more execution time can be saved.
Fig. 8. The correlations between the benchmark dataset $D_1$ and the resampled datasets on: (a) weekday, and (b) weekend.
Fig. 9. Kernel density maps on drop-off points on workday and weekend using the datasets $D_1$, $D_5$ and $D_{10}$. The left column represents workday, and the right column represents weekend.
6. Conclusion

This paper proposes, validates, and analyses a generally applicable data processing framework and its related algorithms to derive micro-mobility patterns by exploiting vehicle availability data. Using e-scooter availability datasets from Zurich, Switzerland, our rigorous analysis shows that data processing indeed has a significant and intricate impact on the mobility patterns derived. This demonstrates the applicability and importance of our research for future researchers and practitioners using micro-mobility vehicle availability data. The framework and recommendation are especially relevant if the data will be used to make evidence-based correspondingly.

Table 2

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Cosine similarity Workday</th>
<th>Cosine similarity Weekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$ and $D_2$</td>
<td>0.983</td>
<td>0.975</td>
</tr>
<tr>
<td>$D_1$ and $D_3$</td>
<td>0.983</td>
<td>0.975</td>
</tr>
<tr>
<td>$D_1$ and $D_4$</td>
<td>0.983</td>
<td>0.974</td>
</tr>
<tr>
<td>$D_1$ and $D_5$</td>
<td>0.982</td>
<td>0.976</td>
</tr>
<tr>
<td>$D_1$ and $D_6$</td>
<td>0.981</td>
<td>0.974</td>
</tr>
<tr>
<td>$D_1$ and $D_7$</td>
<td>0.980</td>
<td>0.974</td>
</tr>
<tr>
<td>$D_1$ and $D_8$</td>
<td>0.980</td>
<td>0.976</td>
</tr>
<tr>
<td>$D_1$ and $D_9$</td>
<td>0.979</td>
<td>0.973</td>
</tr>
<tr>
<td>$D_1$ and $D_{10}$</td>
<td>0.979</td>
<td>0.975</td>
</tr>
</tbody>
</table>

Fig. 10. Basic statistics of three micro-mobility descriptors via boxplot, including (a) duration, (b) distance, and (c) average speed.
transportation policy, such as the central monitoring and evaluation across all e-scooter trial areas in the UK mentioned in the introduction.

In particular, the case study has three major findings that could help future researchers choose an appropriate sampling rate when processing vehicle availability data. First, the sampling rate does not show remarkable influences on micro-mobility in the temporal

Fig. 11. Statistical distribution of three micro-mobility descriptor using datasets $D_1, D_2, D_4, D_6, D_8$, and $D_{10}$, which are displayed from top to bottom in turn.
and spatial descriptors. It implies that if the aim of the research is to uncover the temporal variations of the number e-scooter trips or the spatial distribution of aggregate travel demand on e-scooter, it might be not necessary to collect or use the dataset with high sampling rate. Second, the statistical analysis is conducted to quantify the effects based on three typical mobility descriptors, including trip duration, distance and speed. On one hand, the descriptive statistics of the three descriptors are calculated via boxplot, including median, percentiles, etc. It is found that trip duration is overestimated and trip speed is underestimated while improving the temporal sampling interval. The trip distance is not influenced significantly by downsampling of datasets. On the other hand, the statistical characteristics of the descriptors are investigated by plotting their probability density distribution. It is observed that the three descriptors duration, distance and speed display different statistical characteristics, which can be fitted by power law distribution, exponential function and Weibull distribution appropriately respectively. By comparing the changes of the exponents for the statistical distribution models between benchmark dataset and the resampled datasets, it is indicated that the sampling rate impacts the micro-mobility descriptors, especially on duration and speed. Third, the effect of sampling rate is measured in terms of identifying trips from

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Duration</th>
<th>Distance</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>R²</td>
<td>β</td>
</tr>
<tr>
<td>D₁</td>
<td>2.681</td>
<td>0.945</td>
<td>0.107</td>
</tr>
<tr>
<td>D₂</td>
<td>2.796</td>
<td>0.938</td>
<td>0.170</td>
</tr>
<tr>
<td>D₃</td>
<td>2.815</td>
<td>0.943</td>
<td>0.167</td>
</tr>
<tr>
<td>D₄</td>
<td>3.016</td>
<td>0.956</td>
<td>0.164</td>
</tr>
<tr>
<td>D₅</td>
<td>3.187</td>
<td>0.970</td>
<td>0.159</td>
</tr>
<tr>
<td>D₆</td>
<td>3.38</td>
<td>0.958</td>
<td>0.155</td>
</tr>
<tr>
<td>D₇</td>
<td>3.501</td>
<td>0.977</td>
<td>0.142</td>
</tr>
<tr>
<td>D₈</td>
<td>3.663</td>
<td>0.964</td>
<td>0.141</td>
</tr>
<tr>
<td>D₉</td>
<td>3.865</td>
<td>0.967</td>
<td>0.141</td>
</tr>
<tr>
<td>D₁₀</td>
<td>3.945</td>
<td>0.973</td>
<td>0.141</td>
</tr>
</tbody>
</table>

Table 4
Similarity measurement on probability distributions between the benchmark dataset D₁ and each resampled dataset by calculating two types of distance.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Duration</th>
<th>Distance</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HD</td>
<td>BD</td>
<td>HD</td>
</tr>
<tr>
<td>D₁ and D₂</td>
<td>0.014</td>
<td>4.107</td>
<td>0.010</td>
</tr>
<tr>
<td>D₁ and D₃</td>
<td>0.018</td>
<td>4.115</td>
<td>0.009</td>
</tr>
<tr>
<td>D₁ and D₄</td>
<td>0.035</td>
<td>4.170</td>
<td>0.009</td>
</tr>
<tr>
<td>D₁ and D₅</td>
<td>0.052</td>
<td>4.272</td>
<td>0.008</td>
</tr>
<tr>
<td>D₁ and D₆</td>
<td>0.065</td>
<td>4.388</td>
<td>0.007</td>
</tr>
<tr>
<td>D₁ and D₇</td>
<td>0.074</td>
<td>4.493</td>
<td>0.007</td>
</tr>
<tr>
<td>D₁ and D₈</td>
<td>0.081</td>
<td>4.601</td>
<td>0.007</td>
</tr>
<tr>
<td>D₁ and D₉</td>
<td>0.088</td>
<td>4.724</td>
<td>0.007</td>
</tr>
<tr>
<td>D₁ and D₁₀</td>
<td>0.095</td>
<td>4.869</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 5
Comparison on computation efficiency in terms of identifying trips.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>The number of records</th>
<th>Computation time (second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁</td>
<td>3,633,074</td>
<td>910</td>
</tr>
<tr>
<td>D₂</td>
<td>758,689</td>
<td>196</td>
</tr>
<tr>
<td>D₃</td>
<td>516,694</td>
<td>137</td>
</tr>
<tr>
<td>D₄</td>
<td>395,276</td>
<td>113</td>
</tr>
<tr>
<td>D₅</td>
<td>320,885</td>
<td>104</td>
</tr>
<tr>
<td>D₆</td>
<td>263,948</td>
<td>77</td>
</tr>
<tr>
<td>D₇</td>
<td>225,565</td>
<td>62</td>
</tr>
<tr>
<td>D₈</td>
<td>199,030</td>
<td>55</td>
</tr>
<tr>
<td>D₉</td>
<td>178,824</td>
<td>49</td>
</tr>
<tr>
<td>D₁₀</td>
<td>161,673</td>
<td>45</td>
</tr>
</tbody>
</table>

and spatial descriptors. It implies that if the aim of the research is to uncover the temporal variations of the number e-scooter trips or the spatial distribution of aggregate travel demand on e-scooter, it might be not necessary to collect or use the dataset with high sampling rate. Second, the statistical analysis is conducted to quantify the effects based on three typical mobility descriptors, including trip duration, distance and speed. On one hand, the descriptive statistics of the three descriptors are calculated via boxplot, including median, percentiles, etc. It is found that trip duration is overestimated and trip speed is underestimated while improving the temporal sampling interval. The trip distance is not influenced significantly by downsampling of datasets. On the other hand, the statistical characteristics of the descriptors are investigated by plotting their probability density distribution. It is observed that the three descriptors duration, distance and speed display different statistical characteristics, which can be fitted by power law distribution, exponential function and Weibull distribution appropriately respectively. By comparing the changes of the exponents for the statistical distribution models between benchmark dataset and the resampled datasets, it is indicated that the sampling rate impacts the micro-mobility descriptors, especially on duration and speed. Third, the effect of sampling rate is measured in terms of identifying trips from
the perspective of computation efficiency. By taking the data over one day as an example, it is seen that the number of data records and computation time decrease dramatically with the increment of sampling interval.

This research calls for more attention to investigate various issues with the processing of emerging mobility data to ensure its validity for geospatial and transportation research. As valid data builds the foundation for valid research, such studies are of great importance. Despite a thorough investigation, some open questions could be further investigated. For example, instead of using the haversine distance between each pair of pick-up and drop-off point, our near future work will attempt to obtain the approximate trip of the actual trajectory by considering the network constraints and traffic restrictions. Overall, the framework developed in this paper could be readily extended to further investigate issues associated with vehicle availability data or similar issues of other emerging mobility data. All codes of the algorithms are freely accessible online (https://github.com/micromobility-research/VADprocessing).

CRediT authorship contribution statement

Pengxiang Zhao: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. He Haitao: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing. Aoyong Li: Conceptualization, Methodology, Investigation, Writing - review & editing. Ali Mansourian: Conceptualization, Investigation, Writing - review & editing.

Acknowledgements

This research has been supported by the QR Strategic Priorities Fund provided by Research England. The authors are thankful to Roll2Go AG and Lukas Ballo for collecting the raw data used in this research as specified in Section 3, to Daniel J. Reck for providing useful comments on an earlier draft, and to the Editor and three anonymous reviewers for their constructive comments which helped further improve this paper.

Appendix A. Trip identification algorithm

Algorithm 1. Identify trips from GPS records for one vehicle

\[\text{Input} \text{ An array of } n \text{ GPS records } R, \text{ distance threshold } \delta\]

\[\text{Output} \text{ A list of trips } T\]

1. Sort the GPS records chronologically, and get the sorted array \(R' = \{S_1, S_2, \ldots, S_n\}\)
2. while \(i\) belongs to \(0:n\) do
   3. check whether the locations of two consecutive GPS records \(S_i\) and \(S_{i+1}\) are the same
   4. if two locations are the same then \(i = i + 1\)
   5. else calculate the geographic distance dist between \(S_i\) and \(S_{i+1}\), and define a empty list signal\_list to store the drifting points
   6. if \(\text{dist} < \delta\) then indicate GPS signal drifting exits, and store the records \(S_i\) and \(S_{i+1}\) into the list signal\_list, and set \(i = i + 1\)
   7. else check whether the list signal\_list is empty
   8. if signal\_list is empty then the tuple \((S_i, S_{i+1})\) corresponds to one trip, and store it into \(T\)
   9. else calculate the average longitude and latitude of records in signal\_list, and replace the longitude and latitude of \(S_i\) with them; the tuple \((S_i, S_{i+1})\) corresponds to one trip, and store it into \(T\); empty the list signal\_list and set \(i = i + 1\)

References


