Check for updates

IET Intelligent Transport Systems

*Research Article*

IET Journals
The Institution of Engineering and Technology

# Perspectives on stability and mobility of transit passenger's travel behaviour through smart card data

ISSN 1751-956X
Received on 28th March 2019
Revised 28th June 2019
Accepted on 5th August 2019
E-First on 17th September 2019
doi: 10.1049/iet-its.2019.0212
www.ietdl.org

*Zhiyong Cui*[1], *Ying Long*[2] ✉

[1]*Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA*
[2]*School of Architecture and Hang Lung Center for Real Estate, Tsinghua University, Beijing 100084, People's Republic of China*
✉ *E-mail: ylong@tsinghua.edu.cn*

**Abstract:** Existing studies have extensively used spatiotemporal data to discover the mobility patterns of various types of travellers. Smart card data (SCD) collected by the automated fare collection systems can reflect a general view of the mobility pattern of public transit riders. Mobility patterns of transit riders are temporally and spatially dynamic, and therefore difficult to measure. However, few existing studies measure both the mobility and stability of transit riders' travel patterns over a long period of time. To analyse the long-term changes of transit riders' travel behaviour, the authors define a metric for measuring the similarity between SCD, in this study. Also an improved density-based clustering algorithm, simplified smoothed ordering points to identify the clustering structure (SS-OPTICS), to identify transit rider clusters is proposed. Compared to the original OPTICS, SS-OPTICS needs fewer parameters and has better generalisation ability. Further, the generated clusters are categorised according to their features of regularity and occasionality. Based on the generated clusters and categories, fine- and coarse-grained travel pattern transitions of transit riders over four years from 2010 to 2014 are measured. By combining socioeconomic data of Beijing in the year of 2010 and 2014, the interdependence between stability and mobility of transit riders' travel behaviour is also discussed.

## 1 Introduction

The continuum of human spatial immobility–mobility at varying geographic and temporal scales poses fascinating topics and challenges for researchers and governments to make decisions on urban planning and transportation development. Stability and mobility are relevant, since mobility reflects the movement in short-term temporal or small spatial scales, while stability refers to long-term duration of stay or large scale locational consistency. Geographically, people move over scales ranging from a few metres to hundreds of kilometres in metropolitan areas; temporally, they move or stay over scales ranging from a few minutes to many years. Although people's movement seems to be disordered, we can still mining useful patterns for both individuals and a group of residents from various types of data.

The temporal and spatial dynamic mobility pattern of residents have been concerned about for a long time by researchers in the fields of transportation engineering [1], computer science [2], urban planning [3], or even socioeconomics [4]. Along with the development of computer science and geographic information system, many new technologies and new types of data can be utilised to measure people's mobility pattern in large-scale regions, such as call detail records of mobile phone [5], taxicabs' GPS information [6], or even outdoor Wi-Fi signal data. When it comes to the city-wide mobility analysis, smart card data (SCD) collected by automated fare collection (AFC) systems may be a better choice, since AFC system are widely adopted by public transportation operators in most metropolitan areas [7].

AFC systems based on contactless smart cards are available for both city buses and metros to record the details of transaction information when passengers are boarding or alighting. SCD contains fine-grained information not only about passengers' ID (smart cards' ID) and locations of boarding or alighting stations but also transaction time and bus/metro lines. It is a great convenience to utilise SCD to depict passengers' daily, weekly or yearly travel profiles in large-scale regions covered by public transit systems. From an individual perspective, SCD can help record the passenger's travel records, reflect his/her social and economic characteristics, and even forecast his/her routine travel patterns. From a city perspective, SCD acting as a transportation probe can help estimate transportation conditions and provide new materials for intelligent transit systems and urban planning policy. A large number of transit behavioural studies based on SCD have been carried out and gained popularities. However, there are still rooms for improvement in existing studies on SCD-based travel pattern analysis, which can be summarised into three aspects: (i) The SCD as a data source for mobility pattern research has inherent shortcomings; (ii) The existing analysis methods also have drawbacks in terms of model efficiency and complicatedness; (iii) Few studies focus on the long-term evolutionary travel behaviour analysis. These shortcomings are described in detail in the following paragraphs and solved in this study.

Firstly, although SCD has multiple advantages, the shortcomings of SCD are obvious. The anonymous attribute of SCD determines the absence of basic personal information, like age and gender, without which it is hard to measure passengers' socioeconomic characteristics. For analyses spanning a long period of time, it may encounter the problems of changing smart cards and possessing multi-cards of passengers. In addition, some urban transit systems only record when and where passengers board transit buses and neglect the alighting information, leading to the extreme difficulty of inferring the destinations of passengers. Further, due to the increasing volume of rural–urban migration and the transient mobility of the internal migration in megalopolises, the measurement of the stability of the passenger's travel behaviour will be highly influenced. Thus, for the sake of accurately measuring the stability and mobility of the transit passenger's travel behaviour, effective classification or clustering methods should be employed.

However, existing SCD-based travel behaviour analysis methods also have drawbacks. Since SCD normally does not contain the information for labelling the data, such as smart card owner information, classification methods can hardly be applied without labelling data. Hence, most existing studies employed clustering methods. With the help of clustering methods, transit travellers can be clustered into multiple groups, within which the grouped travellers share similar travel patterns. Most clustering methods, such as *K*-means, need to specify the number of clusters,


© The Institution of Engineering and Technology 2019

i.e. the value of $K$, in advance. However, for the SCD clustering task, due to irregularity and variability of the passengers' travel behaviours, it is hard to decide how many clusters should be contained in the SCD set. Some density-based clustering methods with no need to specify the number of clusters, such as the density-based scanning algorithm with noise (DBSCAN) and ordering points to identify the clustering structure (OPTICS), have been applied for SCD clustering. However, these methods normally need extra efforts to pre-set some meta parameters for these models, such as the distance threshold between clusters. Thus, to make the clustering process more efficient and convenient, we propose a simplified smooth OPTICS clustering method to group the SCD in this study.

Much SCD-based work [8–10] focusing on mining transit passenger travel patterns attempted to distinguish the commuting trips, including non-transfer and with-transfer trips, and non-commuting trips. These various types of grouped trips are important to realise many short-term applications, such as travel time prediction and transportation scheduling. Even though existing studies have been conducted to characterise long-term urban dynamics using SCD [11], how do travel patterns change over the years has not been well-studied. The long-term travel pattern changes are capable of reflecting the stability and mobility patterns of transit passengers and revealing the transportation development and the underlying urban evolution.

In this study, we utilise the temporal information of SCD to mine the relationship between the passenger's mobility and stability in different times and frequency scales. To overcome the drawbacks of the SCD, we define a metric for measuring the distance between different SCDs to better describe their similarity. Further, we propose an advanced density-based clustering method to group transit trips into different clusters. Based on the clustering results, we utilise SCD collected from different years to characterise the long-term stability and mobility of transit passengers' travel patterns. To better understand the passenger behaviour in public transportation, we introduce other socioeconomic data into our analysis. Our contributions can be described as follows:

- We define an SCD similarity metric for measuring the difference between passengers' travel behaviours. To better describe the similarity, both the temporal difference and the frequency difference between SCD records are considered.
- We propose a simplified smoothed OPTICS (SS-OPTICS) clustering method to cluster SCDs. Comparing to the classical OPTICS methods, the SS-OPTICS needs less parameters. We discover groups of passengers behaving similarly with respect to their boarding time.
- We cluster the SCD based on different grouping granularities to analyse the long-term mobility and stability of transit passengers. To measure the evolutionary changes in the clustering results, socioeconomic data in different years is also incorporated in our analysis.

The remaining part of this paper is organised as follows. Related work is briefly discussed in Section 2. In Section 3, we proposed the SS-OPTICS algorithm and describe our methodology. Our analysis of mobility and stability is present in Section 4. Section 5 concludes the paper with a summary and a short discussion of future research.

## 2 Related work

The concepts of stability and mobility of travel patterns are opposite but relevant. With the regard of the travel pattern analysis, the mobility tends to reflect the variations of passengers' travel patterns, while the stability characterises the steadiness of patterns over a period of time. Hanson [4] was among the first researchers to focus on stability analysis and stated that analysing individuals' stability also requires analysing their mobility. Through an empirical example centred on the relationship between entrepreneurship and place, he explicitly proposed that considering locational stability requires examining stability and mobility in tandem, since spatiotemporal dynamics involved. Based on this idea, Bagrow and Lin [12] concentrated on detailed substructures and spatiotemporal flows of mobility to show that individual mobility is dominated by small groups of frequently visited, dynamically close locations, forming primary 'habitats' capturing typical daily activity. While many other works [1, 13–15] chose a perspective on large-scale mobility about urban human beings, vehicles or taxis.

To measure residents' stability and mobility in the urban area, SCD in public transit is one of the most widely used data. According to Long *et al.* [16], SCD related research topics can be classified as: (i) data processing and data complementation, such as back-calculation of origin and destination and recognition of trip purpose; (ii) supporting and management of public transit systems; (iii) place-based urban spatial structure and (iv) person-based analysis on the social network and special group of people. Pelletier *et al.* [17] also gave a literature review of SCD usage in public transit and presented three levels of management of SCD: strategic (long-term planning), tactical (service adjustments and network development), and operational (ridership statistics and performance indicators). Zheng *et al.* [2] presented several typical applications based on SCD, like building more accurate route planners. Further, Long *et al.* [3] sought to understand extreme public transit riders in Beijing using both traditional household surveys and SCD. In their work, public transit riders were classified into four groups of different types of extreme transit behaviours to identify the spatiotemporal patterns of these four extreme transit behaviours. Further, Lathia *et al.* [18] discussed personalising transport information services based on SCD. Among their contributions, the authors used clustering methods to prove that the usage of public transportation can vary considerably between individuals. Each passenger's trips were aggregated into a weekday profile describing his temporal habits and hierarchical agglomerative clustering is introduced to discover groups of passengers characterising different travel habits. Contrary to this approach, our weekly profile, presented in Section 3, consisting of hour-grained grids can show more details.

As we investigated, many clustering methods were adopted to process and analyse SCD. To clustering the temporal information, Ifsttar *et al.* [7] constructed temporal passenger profiles based on boarding information and applied a generative model-based clustering approach to discover clusters of passengers. Based on the boarding information, passengers were assigned with 'residential' areas, established through the clustering of socioeconomic data, to inspect how socioeconomic characteristics are distributed over the passenger temporal clusters. To analyse year-to-year changes in public transport passenger behaviour, Briand *et al.* [19] proposed a two-level generative model that applies the Gaussian mixture model to regroup passengers based on the passengers' temporal habits in their public transit usage. A density-based clustering method, DBSCAN [20], which is very similar to OPTICS [21] is used by Ma *et al*. [1]. The authors identified trip chains to detect transit riders' historical travel patterns and apply $K$-Means++ clustering algorithm and the rough-set theory to cluster and classify travel pattern regularities. To detect and update the daily changes in travel patterns, a weighted stop-based DBSCAN is also proposed to reduce computation complexity [22]. To achieve better clustering performance, a two-step clustering method [10] was proposed to cluster transit stations and passengers, respectively. Compared to the approaches presented in these works, we improve the OPTICS algorithm to cut down input parameters and control cluster size. Further, other than, focusing on people's mobility pattern, we utilise SCD to measure the interdependence between stability and mobility in the time dimension.

## 3 Profiling passengers based on SCD

The transit passengers profiling process consists four stages: (i) pre-processing the SCD based on existing studies; (ii) defining the distance (similarity) between different SCD records; (iii) clustering samples of SCD with a proposed simplified smoothed OPTICS


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

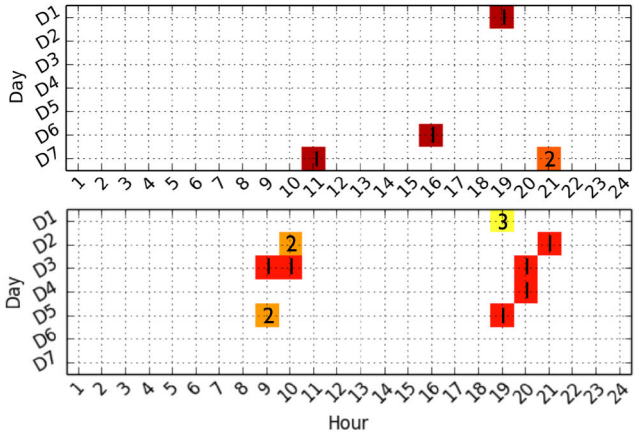


**Fig. 1** *Weekly profiles of two passengers' transaction time. The transaction time (coloured squares) reflects their different travel patterns. The values in coloured squares represent the number of transactions in that specific hour. D1–D5: weekdays, D6: Saturday, D7: Sunday*

**Table 1** Definitions of extreme travellers, according to [3], which are eliminated in the experiments in this study

| Type | Definition |
|---|---|
| early birds (EBs) | First trip < 6 AM, more than two days in five weekdays (60% of weekdays) |
| night owls (NOs) | Last trip > 10 PM, more than two days in five weekdays (60% weekdays) |
| tireless itinerants (TIs) | > = one and a half hours commuting, more than two days in a week |
| recurring itinerants (RIs) | > = 30 trips in weekdays of a week (> = 6 trips per day) |

algorithm; and (iv) classifying the whole SCD records with a *K*-means-like algorithm according to results of the clustering stage.

### 3.1 Dataset description

The SCD collected and issued by Beijing Transit Incorporated contains transit riders' records for both the bus and metro systems. There were two types of AFC systems on Beijing buses: flat fares and distance-based fares, before the beginning of 2015, since when all bus lines became distance-based fare systems. It is a design flaw for the bus smart card system that flat fares system records the transaction (paying) time when checking-in, whereas distance-based fares system records the transaction time when checking-out. For the Beijing metro system, although passengers pay the fare when alighting, the system records the time of both checking-in and checking-out. In this paper, to offset the design flaw, we consider the transaction time as the time for one ride.

We select SCD with shared card IDs from two datasets in 2010 and 2014. Both the selected datasets of 2010 and 2014 last for one week and contain the same smart card IDs with the amount of 1.9 million, representing 1.9 million passengers lived in Beijing at least from 2010 to 2014. We assume each smart card represents an anonymous passenger, without considering the situation of passengers' changing card, which is not common in Beijing. Each record of the SCD consists of (i) smart card ID, (ii) boarding or alighting time, and (iii) station ID of boarding or alighting line. As the time spans of SCD in 2010 and 2014 both cover one week, we estimate each passenger's trip activities using a 'weekly profile', a vector contains 168 ($7 \times 24$) variables describing the distribution of the trip activities. Each variable in the vector represents the number of smart card's transaction time over each hour in each day of the week. Fig. 1 illustrates weekly profiles of passengers' transaction time.

### 3.2 Data pre-processing

Before analysing the transit behaviours of the passengers in Beijing, we separate the whole passengers into two groups: extreme travellers and non-extreme travellers, according to an existing study [3]. Four types of extreme travellers are defined based on their behaviours in weekdays, by setting several validated thresholds and combining empirical knowledge of Beijing as depicted in Table 1. For example, since most people's working hours start at 8:30 or 9:00 am in Beijing, public transit boarding time before 6:00 am would be considered as an unusually early situation [3]. As we evaluated, those types of extreme travellers only account for a small proportion (less than 5%) of the whole passengers. Since the extreme travellers have clear definitions and can be easily filtered out from the raw data, the data of extreme travellers is eliminated from the dataset for clustering and analysed separately in the experiments in our study.

### 3.3 Definition of distance between smart card records

After counting the number of transaction time of each smart card record, the feature of each record forms a travel record vector with 168 ($24 \times 7$) elements, $\boldsymbol{v} = [v_0, \ldots, v_{167}]$, representing the number of smart card transactions in each hour of each day in the study period (one week). Each element of $\boldsymbol{v}$ is a non-negative integer that $\boldsymbol{v}_i \in \mathbb{Z}$. There are some classic distance-measurement methods to measure the similarity of different records, such as Euclidean distance, Manhattan distance, cosine distance, and cross correlation distance (CCD). However, since each transaction vector not only records the number of transactions but also contains temporal attributes, those classical distance formulas are not capable of comprehensively measuring the difference, i.e. the distance, between smart card transactions.

The Euclidean distance can only measure the straight-line distance between two points, represented by two vectors with 168 dimensions in this study. Since the order of dimensions in the Euclidean space does not influence the distance, the Euclidean distance inherently misses out the temporal information between the SCD records. Similar to the Euclidean distance, the Manhattan distance measures the grid-based distance between two points and fails to consider the temporal information. The cosine distance, i.e. the cosine similarity, measures the similarity between two non-zero vectors of an inner product space and output the cosine of the angle between those two vectors. The cosine distance, commonly used in high-dimensional positive spaces, has already been used in SCD analysis [23]. However, since the travel record vectors are mostly sparse, the calculation of cosine distance conducting inner products on the two compared vectors will significantly cancel out the difference between the elements in the two vectors, if one of the elements in a vector is zero. For example, for the two travel record vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, if $v_i = 0$ and $u_i \neq 0$, the cosine distance of $\boldsymbol{u}$ and $\boldsymbol{v}$ at the ith position $\propto v_i * u_i = 0$, which still neglect the travel frequency difference at the ith time point. The CCD has also been used to measure similarity/distance between two sequences/vectors by shifting one sequence to find a maximum correlation with another sequence [24]. However, due to the shifting mechanism, the calculation process of CCD almost gets rid of the relative position information of elements in two sequences. That means


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

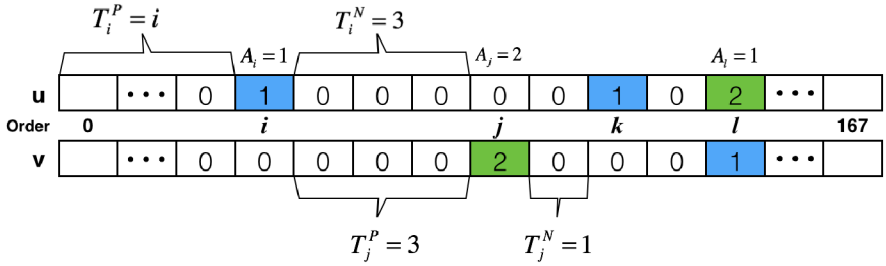


**Fig. 2** *Example of the distance between two vectors, **u** and **v***

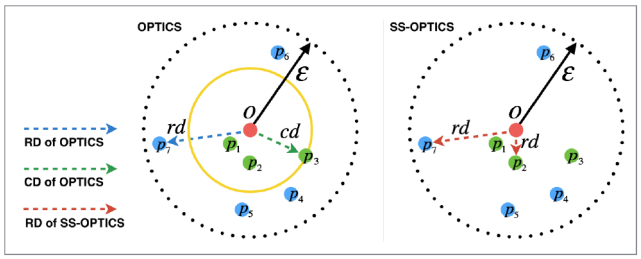


**Fig. 3** *Cd of OPTICS and the rd of both OPTICS and SS-OPTICS. MinPts = 4*

CCD cannot capture the difference of occurrence time of smart card transactions, and thus, CCD is not utilised in this study.

To solve this problem, we propose a new distance metric between two SCD transaction record vectors, defined as *transaction distance* (TD), to measure the difference between two transit passengers' travel patterns. Since most passengers normally do not take transit very frequent, most of the elements in a travel record vector are zeros. In this case, the non-zero values in the travel record vector greatly impact the difference between different vectors. We define the TD between the two vectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, by considering both the time difference and the frequency difference. Thus, the proposed TD consists of two parts, the riding time interval ($T_i$) and the absolute riding frequency difference ($A_i$), for each element $i$.

The absolute riding frequency difference is defined as $A_i = |u_i - v_i|$. As for the riding time interval $T_i$, if one of $u_i$ and $v_i$ equals to 0, $T_i$ equals the smaller value of $T_i^{\mathrm{P}}$ and $T_i^{\mathrm{N}}$, namely $T_i = \min\left\{T_i^{\mathrm{P}}, T_i^{\mathrm{N}}\right\}$. Here, taking $u_i \neq 0$ for example, $T_i^{\mathrm{P}}$ represents the time interval between the current element $u_i$ and the nearest *previous* non-zero element in vector $\boldsymbol{v}$. Also correspondingly, $T_i^{\mathrm{N}}$ represents the time interval between the current element and the nearest *next* non-zero element in vector $\boldsymbol{v}$. If $u_i$ and $v_i$ both equal to or do not equal to 0, $T_i = 0$. Fig. 2 shows an example of how to compute the transaction distance that $T_j = \min\{T_j^{\mathrm{P}}, T_j^{\mathrm{N}}\} = 1$ and $T_l = 0$. If a non-zero component in one vector cannot find a previous or next non-zero component in the other vector, like the situation of $u_i$ in Fig. 2, its $T_i^{\mathrm{P}}$ equals $\min\{i, 167 - i\}$.

Then, the transaction distance between vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ can be represented as

$$\mathrm{TD} = \sum_{i=0}^{167} \min\{T_i^{\mathrm{P}}, T_i^{\mathrm{N}}\} + k * \left|u_i - v_i\right|, \ \mathrm{s.t.} \quad u_i \neq v_i \qquad (1)$$

where $k$ is a parameter to balance the weights of $T$ and $A$. It is suitable for setting the value of $k$ ranging from 0 to 3, as we tested in the clustering section.

### 3.4 Clustering samples of SCD records

After defining the distance between vectors of smart card records, we cluster the vectors to identify the travel patterns of public transit riders in Beijing. To accurately cluster the travellers, a suitable algorithm is needed. Although $K$-Means algorithms or other centroid-based clustering models are very efficient, they need to nominate the number of clusters ($K$) before running of the algorithm. Even though iteratively setting $K$ as different numbers and evaluating the performance may help to identify the proper value of $K$, dealing with a high number of observations during this iterative process is still a big problem for $K$-Means. As the travel record vector has a high dimension (168) in this study, grid-based clustering methods are also not suitable for this problem. A new density-based clustering algorithm, clustering by fast search and find of density peaks [25], is also tested. However, it can only identify 4 or 5 obvious clusters. Thus, to solve the aforementioned problems, we propose an improved density-based clustering algorithm based on OPTICS [21], which is suitable for clustering the data based on the aforementioned TDs. We name it as the simplified smoothed OPTICS (SS-OPTICS).

*3.4.1 Simplify:* The original OPTICS algorithm has two key concepts, *core-distance* and *reachability-distance*.

*Definition-1: $\varepsilon$ -Neighbourhood*: Let $p$ be an object from a dataset D. The $\varepsilon$-neighbourhood set of a point $p$ is defined by $N_\varepsilon(p) = \{x \in D | dist(p, x) \leq \varepsilon\}$.

*Definition-2: : Core-distance (cd)*: Let $p \in D$, let $\varepsilon$ be a distance value, let $N_\varepsilon(p)$ be the $\varepsilon$ -neighbourhood of $p$, let MinPts be a natural number and let MinPts-distance*(p)* be the distance from $p$ to its MinPts neighbour. Then, the core-distance of $p$ is defined as $\text{core-distance}_{\varepsilon,\text{MinPts}}(p) =$

$$\begin{cases} \text{UNDEFINED}, & \text{if Card}(N_\varepsilon(p)) < \text{MinPts} \\ \text{MinPts} - \text{distance}(p), & \text{otherwise} \end{cases}$$

*Definition-3: : Reachability-distance (rd)*: Let $p, o \in D$, let $N_\varepsilon(o)$ be the $\varepsilon$ -neighbourhood of $o$, let MinPts be a natural number. Then, the reachability-distance of $p$ with respect to $o$ is defined as $\text{reachability-distance}_{\varepsilon,\text{MinPts}}(p, o) =$

$$\begin{cases} \text{UNDEFINED}, & \text{if } |N_\varepsilon(o)| < \text{MinPts} \\ \max\ (\text{core} - \text{distance}(o),\ \text{distance}(o, p)), & \text{otherwise} \end{cases}$$

where $\varepsilon$ and MinPts are two input parameters of the original OPTICS algorithm. According to OPTICS's definitions, the green points covered by the yellow circle in Fig. 3 share the same reachability-distance (*rd*), which equals to the core-distance of point $o$ (*cd*). Although the green points, $p_1$, $p_2$, and $p_3$, have the same *rd*, their actual reachable distances from point $o$ are different ($rd'_{p_1} < rd'_{p_2} < rd'_{p_3}$).

The main ideas of OPTICS can be described as (i) reachability distance represents density and (ii) reachability-distance determines the points' output order, which determines clusters. Based on these ideas, we can find a design flaw of OPTICS that the output order of $p_1$, $p_2$, and $p_3$ in the left example of Fig. 3 maybe disordered due to their same *rd*s. Thus, we design an improved OPTICS algorithm by abandoning the concept of core-distance and define a new concept of reachability-distance (RD) as follows (Fig. 4).

*Definition-4: : New Reachability-distance (RD)*: Let $p, o \in D$, let $N_\varepsilon(o)$ be the $\varepsilon$ -neighbourhood of $o$. The reachability-distance of $p$ with respect to $o$ is defined as $\text{reachability-distance}_\varepsilon(p, o) =$


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\begin{cases} \text{UNDEFINED}, & \text{if } |N_\varepsilon(o)| = 0 \\ \text{distance}(o,p) \quad \text{s.t.} \quad p \in N_\varepsilon(o), & \text{otherwise} \end{cases}$$

*3.4.2 Smooth:* The 2D plot based on the ordered points' reachability distance can help us distinguish the clusters. As the denser the points gather, the lower reachability-distances the points get, the 'valley' shapes in the reachability distance curve represent clusters with high density. In Fig. 5, the blue line is the *rd* curve of OPTICS, the green line is the *RD* curve of SS-OPTICS. We notice that, although the average value of SS-OPTICS's RD is obviously less than OPTICS's, their curves are extremely similar.

The red line is the smoothed RD of SS-OPTICS, RD′, in Fig. 5. We smooth the RD curve with two aims. One is to make it easier to identify the valley-shaped clusters, and the other is to control the size of a cluster. By using the mean filtering method to smooth the RD curve, we can achieve the two goals with only one parameter, window size (S). Each value of the smoothed RD curve, $\text{RD}'_i$, is the mean of RD value of points within the window

$$\text{RD}'_i = \left( \sum_{j=i-n}^{j=i+n} \text{RD}_j \right) \Big/ S, \quad \text{s.t.} \quad n = \frac{S-1}{2} \tag{2}$$

Since RD′ has been filtered by a $S$ sized window, it should be noticed that the boundary of the valley-shaped cluster has a bias to the left, and the offset is $(S-1)/2$. After the mean filtering, the valley (cluster) of the RD curve, whose number of the points in this cluster is less than $(S-1)/2$, will nearly be filled up. Thus, the cluster size is controlled to be larger than $(S-1)/2$.

Comparing to OPTICS, SS-OPTICS needs one parameter ($\varepsilon$) and OPTICS needs two ($\varepsilon$ and MinPts). The time complexity of SS-OPTICS is $O(n^2)$, same as OPTICS. Meanwhile, both the algorithms are not sensitive to the value of the parameters. The $\varepsilon$ is set to be 100. In addition, the SS-OPTICS is easier to control the cluster size and define the cluster boundary by setting the window

**Data:** D (Unprocessed Dataset), $\varepsilon$
**Result:** OrderedPoints
initialization;
**while** $D \neq Null$ **do**
  $Point = D.pop()$;
  $OrderedPoints.append(Point)$;
  $P\_neighbors = point.neighbor(\varepsilon)$;
  **if** $P\_neighbors \neq Null$ **then**
    $OrderSeeds$ = [];
    $OrderSeeds.updateRD(Point, P\_neighbors)$;
    **while** $OrderSeeds$ **do**
      $OrderSeeds.sort(key = RD)$;
      $Seed = OrderSeeds.pop()$;
      $OrderedPoints.append(Seed)$;
      $S\_neighbors = Seed.neighbor(\varepsilon)$;
      **if** $S\_neighbors \neq Null$ **then**
        $OrderSeeds.updateRD(Seed, S\_neighbors)$

**Fig. 4** *Algorithm 1: Getting ordered points by OPTICS*

size $S$. Since the window size $S$ only affects the boundary of clusters, it does not influence the overall clustering performance. Thus, after experimental testing, the window size ($S$) is set to be 40 in this study. Finally, we iteratively cluster several random samples of SCD, containing 20,000 entries in each sample, and identify 33 clusters for the next stage to classify the whole dataset. The sensitivity analysis on sample size shows that when the sample size is over 20,000, the clustering results nearly converge, and thus, the sample size is set as 20,000.

According to the transaction time distribution of the 33 clusters, they can be classified into four big categories obviously as shown in Fig. 6. The four categories can be described as: one-day trips, two-day trips, multi-days trips, and commuting trips. The one-day trips containing 7 clusters (9–15) are distributed in one day of the week from Monday to Sunday. The transaction time of two-day trips (cluster 1–8, 16 and 18–23) is distributed mainly in two days of the week, while the transaction time of multi-day trips (cluster 24–27, 29 and 31–33) is dispersed in different days (at least 3 days) of the week. The commuting trips (clusters 17, 28 and 30) are mainly characterised with regular morning and evening peaks during the week.

### 3.5 Classifying SCD records based on SS-OPTICS results

The set of the 33 clusters acquired by SS-OPTICS is denoted as $C = [C_1, \ldots, C_{33}]$. Each $C_i$ in $C$ is a vector containing 168 components, like the vector of smart card's transaction time. Each component $(c_j)$ of cluster $C_i$ is the frequency of passengers' travel behaviour in the $(j\%24)$th hour of the $(j - j\%24/7)$ day of the week. We also add a cluster to $C$ as the 34th cluster, whose components are all zero, to classify some noise points. Hence, taking the clusters' features as the centroids of clusters, $C = [c_{ij}]_{34\times168}$, all the SCD records can be grouped into 34 clusters.

According to the acquired data, two necessary aspects of information for clustering the whole dataset are obtained, i.e. (i) the cluster number $K = 34$ and (ii) the feature of each cluster $C_i$. With the two aspects of information, classifying the whole dataset can be easily fulfilled by utilising the simple and efficient $K$-means algorithm. For each SCD vector, $\boldsymbol{v} = [v_0, \ldots, v_{167}]$, it belongs to Cluster $C_i$ satisfying

$$i = \arg\max_i \sum_{j=1}^{168} v_j \times c_{ij} \tag{3}$$

Then, the cluster $C_i$ is updated such that $c_{ij} =$

$$\begin{cases} \dfrac{n \times c_{ij} + v_j}{n + \| \boldsymbol{v} \|_0}, & \text{if } v_j \neq 0 \\ \dfrac{n \times c_{ij}}{n + \| \boldsymbol{v} \|_0}, & \text{otherwise} \end{cases}$$

where $n$ is the total number of transactions in $C_i$. After iteratively grouping all the SCD records, the clustering results are acquired and analysed in the following section.

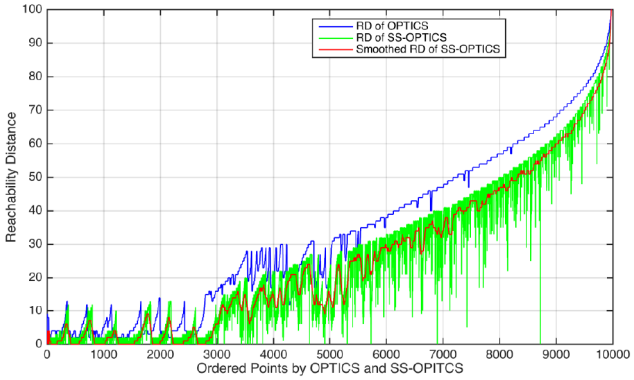


**Fig. 5** *RD curves of OPTICS and SS-OPTICS, $\varepsilon = 100$ and $S = 41$*


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

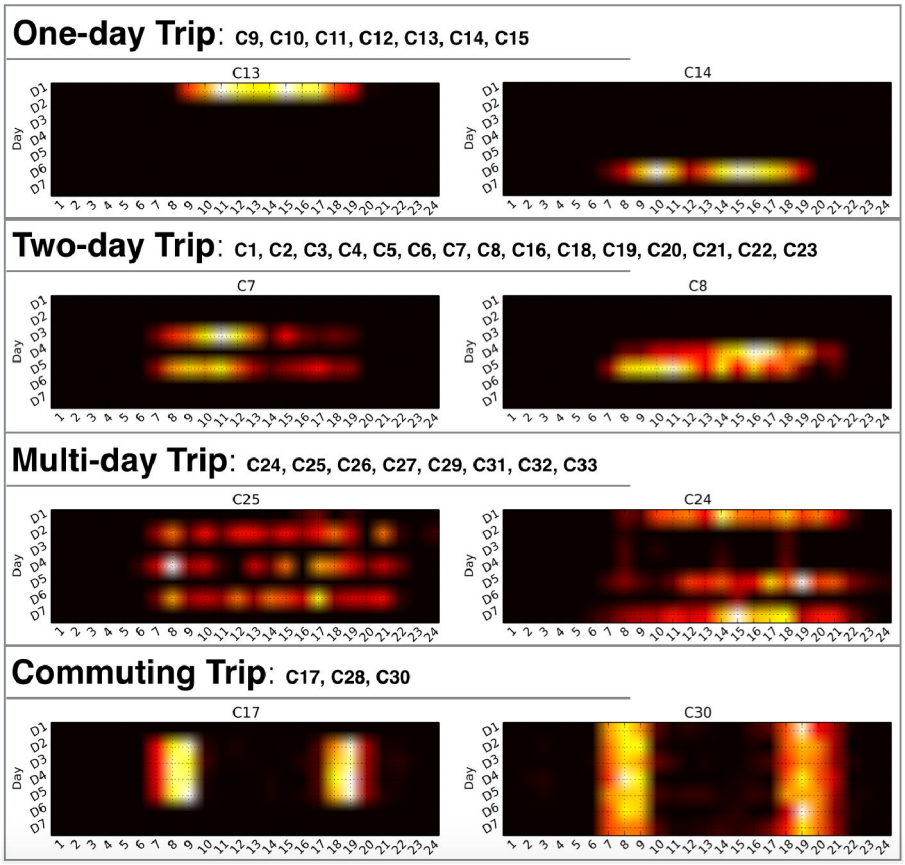


**Fig. 6** *4 obvious categories of the heatmap of the 33 clusters. D1–D5: weekdays, D6 and D7: weekends*

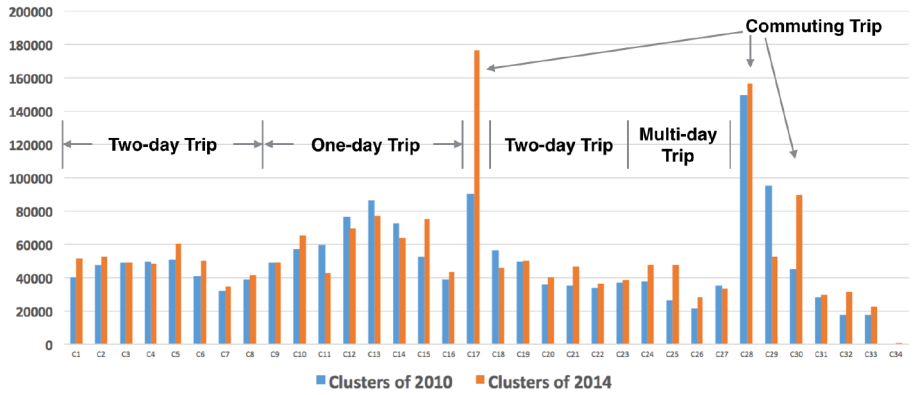


**Fig. 7** *Number of smart cards in each of the 34 clusters in 2010 and 2014*

### *3.6 Clustering results and analysis*

After the clusters were created for all data combined, the data sets of 2010 and 2014 can be classified accordingly. Then, we can get the numbers of the smart card records in each of the 34 clusters in the years of 2010 and 2014. Fig. 7 demonstrates the number of cards in each cluster. It can be noticed that the numbers of trips in one-day, two-day and multi-day trips do not vary much over the four years from 2010 to 2014. The number of one-day trips, around 60,000 in each cluster, is a little more than that of two-day trips, around 50,000. Comparing to the clusters belonging to the multi-day trip group, the one-day and two-day trips occupy the most in the total trips. It reveals that more passengers in Beijing choose to use public transit occasionally, mainly in one day or two days. The number of multi-day trips (about 30,000 in each cluster) is the least. This makes sense that fewer passengers would ride public transit vehicles or metros on multiple days in a week if they are not commuters. It should be noted that although special case, like misclassification, may exist in the clustering results, the overall analysis can hardly be influenced with the help of the huge size of the data set.

Almost, all the towering bars in Fig. 7 belong to the commuting trip group. This group, whose travel patterns are shown in the commuting trip group in Fig. 6, clearly represents the main members of public transit riders, namely the commuters who take a home-to-work trip in the morning and go back home in the evening every weekday. One interesting result is that the number of passengers belonging to the commuting trip group nearly doubled from 2010 to 2014. There are two potential reasons leading to this huge increase in commuting trips in Beijing. One is that public transit became more convenient from 2010 to 2014. As we investigated, during this time period, Beijing metro constructed 8 more lines into 15 lines in total and the total metro length increased rapidly from 228 to 465 km. The other reason is the ground transportation in Beijing became more congested and forced some people to choose public transit, since the total number of private vehicles in Beijing increased from 2.9 million in 2010 to 4.3 million in 2014.


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 2** Social economics factors in Beijing

| Year | Population | Population density | Private vehicles | Bus volume | Metro volume |
|---|---|---|---|---|---|
| 2010 | 17.55 mil. | 1224 person/km$^2$ | 2.97 mil. | 5.165 bil. | 1.423 bil. |
| 2014 | 21.15 mil. | 1498 person/km$^2$ | 4.25 mil. | 4.843 bil. | 3.205 bil. |

**Table 3** Transition matrix of extreme travellers

| | | | | 2014 | | |
|---|---|---|---|---|---|---|
| 2010 | EB | NO | TI | RI | NE | SUM |
| EB | 1286 | 206 | 535 | 82 | 7605 | 9714 |
| NO | 299 | 2550 | 2200 | 153 | 30,006 | 35,208 |
| TI | 376 | 996 | 9488 | 182 | 48,406 | 59,448 |
| RI | 93 | 198 | 677 | 275 | 7351 | 8594 |
| NE | 8780 | 26,357 | 82,630 | 3977 | 1,646,118 | 1,767,862 |
| SUM | 10,834 | 30,307 | 95,530 | 4669 | 1,739,486 | 1,880,826 |

EB: Early Birds, NO: Night Owls, TI: Tireless Itinerants, RI: Recurring Itinerants and NE: Non-Extreme Travellers.

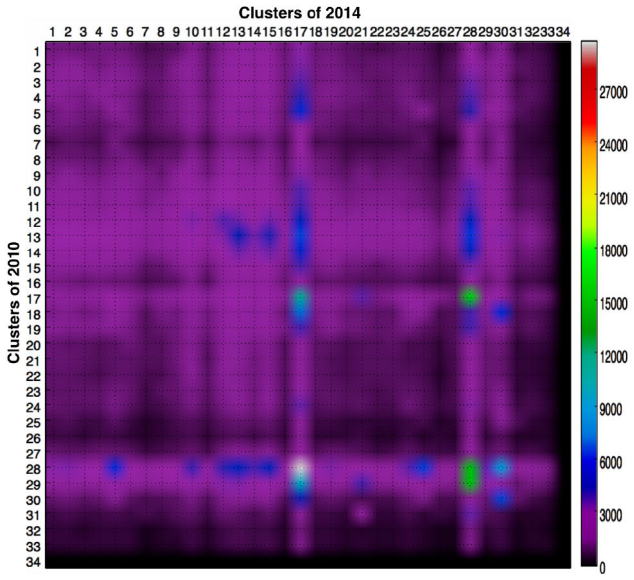


**Fig. 8** *Heatmap of the 34 clusters' transition matrix*

## 4 Mobility and stability analysis

Mobility and stability patterns of people living in metropolitan areas are really hard to measure due to the huge number of residents and incomplete methods to probe all the population. As mentioned by many studies [1, 7, 17], utilising SCD collected by AFC system is a nearly ideal solution of this problem, since public transit is used by a large proportion of urban residents and AFC system can record their travel details. However, we still need to consider the influence of many other factors, including residents age distribution, social scale, per capita income, type of job, city size and so on, to analyse transit passengers' travel behaviours. Since the datasets of 2010 and 2014 are selected according to the same smart card IDs, the mobility and stability of fixed passengers can be reflected by the changes of their travel patterns between 2010 and 2014. Passengers' travel patterns are represented by the recorded transaction time when they using public transit. In this section, we analyse passenger's mobility and stability pattern based on temporal information combining some background socioeconomic factors listed in Table 2.

### *4.1 Extreme travellers analysis*

According to the classification criteria proposed in Table 1, the transition matrix of the four types of extreme travellers (EB, NO, TI, RI) from 2010 to 2014 is generated and shown in Table 3. The numbers of extreme travellers in 2010 (141,340) and 2014 (112,964) are both very small compared to that of non-extreme travellers. In addition, 84% of the extreme travellers in 2010 converted into non-extreme travellers in 2014, which means the stability of extreme travellers' live pattern cannot last for a long time.

However, among the four types of extreme travel patterns, it still can be found that the most passengers in the EB, NO, and TI groups in 2010, with the amounts of 1286, 2250 and 9488, staying in the same groups in 2014. Hence, that means passenger with an extreme travel pattern is more likely to keep the original travel pattern other than to convert into other extreme patterns. It also meets the findings of the previous work [3] that most of EB, NO, and TI are full-time workers, implying full-time worker will less likely change their jobs (also travel pattern) compared to the unemployed.

### *4.2 Non-extreme travellers analysis*

*4.2.1 Fine-grained analysis:* By acquiring the numbers of passengers of 34 clusters in 2010 and 2014, the transition (mobility) matrix of these clusters is calculated and demonstrated by a heatmap shown in Fig. 8. In this heatmap, the brighter the grid is, the more passengers belong to this grid. We can easily catch sight of brighter parts (green, blue and white parts) and find them mainly distributed in cluster 17, cluster 28 and cluster 30 in both 2010 and 2014, which belong to the commuting trip category.

Especially for the green and white grids ($C_{17\to17}$, $C_{17\to28}$, $C_{28\to17}$ and $C_{28\to28}$), the numbers of trips in these grids are several times larger than that of other grids. This reflects the travel pattern stability of the passengers belonging to the commuting trip


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

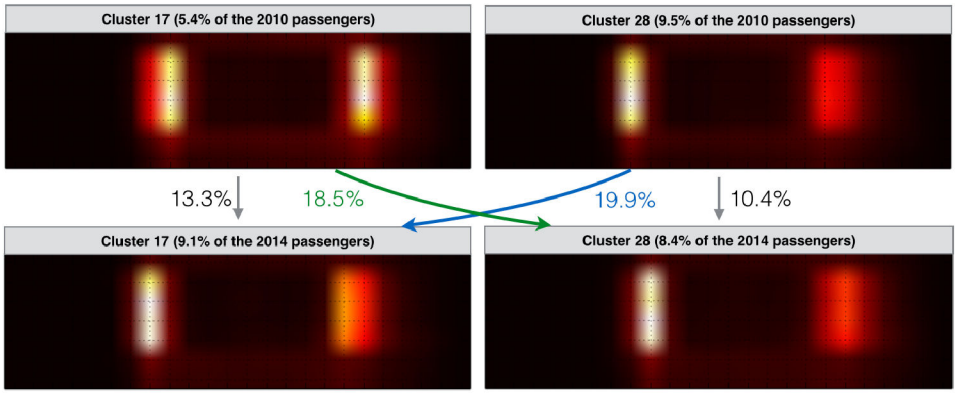


**Fig. 9** *Heatmaps of the mutual transition between cluster 17 and cluster 28 in both 2010 and 2014*

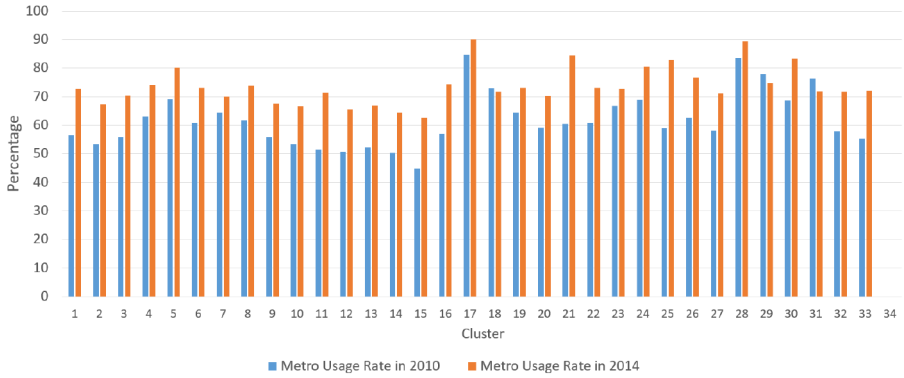


**Fig. 10** *Percentage of the passengers who take the metro at least once a week in each cluster*

**Table 4** Transition matrix of non-extreme travellers

| | | | 2014 | | |
|---|---|---|---|---|---|
| 2010 | O | T | M | C | SUM |
| O | 119,270 | 193,290 | 84,311 | 31,864 | 428,735 |
| T | 164,817 | 298,667 | 142,436 | 48,043 | 653,963 |
| M | 64,449 | 142,399 | 76,769 | 20,038 | 303,655 |
| C | 36,642 | 79,266 | 47,757 | 11,812 | 175,477 |
| SUM | 385,178 | 713,622 | 351,273 | 111,757 | 1,561,830 |

O: One-day trip, T: Two-day trip, M: Multi-day Trip and C: Commuting Trip.

category. These four grids' weekly profiles are demonstrated by heatmaps in Fig. 9.

Although their morning and evening peak hours have a deviation of one hour, the stability can be reflected by the almost same trip occurrence time distribution and the same time intervals between morning and evening peak hours. Their temporal profiles also present most commuting trips of passengers in Beijing are distributed mainly from Tuesday to Friday. It is interesting to explore why commuting passengers tend to ride public transit on weekdays except for Monday. A possible explanation is the Monday morning syndrome, which means some people feel even more tired out on Monday than on Friday after the relaxation over the weekend.

There are also some blue grids distributed in the one-day trips region (cluster 9–15). The heatmap shows the mutual transitions between one-day trip category (cluster 9–15) and commuting trip category (cluster 17, 28) happen a lot. Passengers in the group of one-day trip category are regarded as the ones using public transit occasionally. This transition shows passengers change their public transit usage patterns from occasional to regular on weekdays. This situation can be the result of many reasons, like changing jobs or working locations, earning enough money to buy a car, or taking metro to work instead of driving. Fig. 10 shows the percentage of passengers who rode a metro at least once a week in each cluster. The average percentage in 2014 is apparently higher than that of 2010. Further, as shown in the figure, the percentages of passengers in the commuting clusters (17, 28, and 30) reach the peaks in both lines of 2010 and 2014. This means commuting passengers may be the most stable group who are most willing to transit by the metro.

*4.2.2 Coarse-grained analysis:* The transition matrix of the four groups of non-extreme travellers is also counted and shown in Table 4. Each component of the transition matrix demonstrates the number of passengers transition from one group to another. Analysing the transition between different groups provides a new perspective to analyse passenger's mobility and stability. However, only with SCD, we cannot prove our conjectures. Hence, to better understand the mobility and stability of passengers, we combine socioeconomic statistics data of Beijing in both 2010 and 2014 [26, 27], shown in Table 2. From 2010 to 2014, the population of Beijing increased by 3.6 million and the population density in the urban area rose from 1224 to 1498 persons per square kilometre. Along with the growth of population, the total number of private vehicles in Beijing increased from 2.97 million to 4.25 million. All these factors show that Beijing became more crowded in the urban area and more vehicles led to more congested ground transportation after 2010. As for the transition matrix, the ratios of components in each row of the transition matrix are very close (approximately O:T:M:C = 6:14:7:2), implying the overall travel patterns of passengers in Beijing did not change much from 2010 to 2014. Although the population and the number of vehicles increased a lot in Beijing, the travel patterns of public transit riders tend to be stable. However, as Table 2 indicates, the total volume of


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

passengers riding metros doubled during the four years, while the volume of passengers taking buses decreased a little bit. This unusual decline might be the result of the rapid construction of the Beijing Metro System targeting at mitigating congestion brought by the increasing population and usage of private vehicles. However, as revealed by the transition matrix that the whole non-extreme travellers nearly keeps the same travel patterns. That means if the passengers' travel demand keeps at a similar level, constructing new metro lines may not be able to fundamentally solve the congestion problem.

### *4.3 Discussion on mobility and stability*

Analysing SCD from different temporal scales can provide different points of view to understand the mobility and stability of transit passenger's travel behaviour. The mobility and stability are relevant and a passenger's weekly travel records show his/her short-term mobility patterns, yet the change of whole passengers' mobility patterns over years may imply the unchangeable of their lifestyle or social status. Along with the increase of population, transit availability, and urban size in Beijing, inhabitant's travel pattern changes a lot, but the distributions of different types of trips nearly keep the same. This reveals that individuals' short-term mobility integrates together and forms the population's long-term stability. One interesting phenomenon is that the total mileage of Beijing metro doubled from 2010 to 2014, during which the number of commuting trips also nearly doubled, as shown in Fig. 7. This cannot be the coincidence, since travel behaviours of inhabitants should be largely determined by the general environments, public infrastructure, and services of the city. The aforementioned fine-grained and coarse-grained comparisons of passengers' transit profiles between 2010 and 2014 both highlight trip category transitions between the 34 clusters and the four groups, which presents the distinctive long-term travel pattern dynamics.

## 5 Conclusions

SCD provide us with new perspectives to observe the operation of our cities. In this paper, we analyse the temporal travel pattern of transit passengers in Beijing by clustering the SCD. To better analyse the SCD, we define a metric, i.e. TD, to measure the similarity or difference between passengers' travel patterns by considering both time difference and frequency difference between SCD records. We also propose a simplified smoothed OPTICS clustering method to cluster SCD. Comparing to the classical OPTICS methods, the SS-OPTICS needs fewer parameters and generates better clustering performance. We cluster the SCD based on different grouping granularities to analyse the long-term mobility and stability of transit passengers. By combining some socioeconomic data, we present several analyses about residents' temporal mobility and stability to elucidate the interdependence between mobility and stability of transit passengers' travel patterns. Extreme travellers are most vulnerable that the stability of extreme travellers' life pattern cannot last for a long time. According to clustering outcomes and our analyses, non-extreme travellers' high mobility is shown by the transition between different fine-grained clusters. However, the stability of their travel patterns is also obvious based on coarse-grained travel pattern categorisation.

In the future study, the proposed TD is very suitable for measuring the similarity of time series with specific physical meanings, like the passenger's travel record sequences in this study. Since the proposed SS-OPTICS algorithm can generate optimal clustering results, it can also be applied in transit analysis related applications. In addition, several improvements can be made based on the work presented herein. Firstly, the accuracy of SCD can be enhanced in the future by adopting robust methods to mitigate the deviations of boarding and alighting time. Secondly, the proposed SS-OPTICS algorithm can be improved aiming to find a better way to define the boundaries of clusters. Thirdly, more fine-grained socioeconomic and geospatial data can be incorporated in the analysis.

## 6 Acknowledgment

This work was supported by the National Natural Science Foundation of China (no. 51408039).

## 7 References

[1] Ma, X., Wu, Y.J., Wang, Y., *et al.*: 'Mining smart card data for transit riders' travel patterns', *Transp. Res. C, Emerg. Technol.*, 2013, **36**, (0), pp. 1–12
[2] Zheng, Y., Capra, L., Wolfson, O., *et al.*: 'Urban computing: concepts, methodologies, and applications', *ACM Trans. Intell. Syst. Technol.*, 2014, **5**, (3), pp. 222–235
[3] Long, Y., Liu, X., Zhou, J., *et al.*: 'Early birds, night owls, and tireless/recurring itinerants: an exploratory analysis of extreme transit behaviors in Beijing, China', *Habitat. Int.*, 2016, **57**, pp. 223–232
[4] Hanson, S.: 'Perspectives on the geographic stability and mobility of people in cities', *Proc. Natl. Acad. Sci.*, 2005, **102**, (43), pp. 15301–15306
[5] Kang, C., Sobolevsky, S., Liu, Y., *et al.*: 'Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages'. UrbComp '13 Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, IL, USA, 2013
[6] Peng, C., Jin, X., Wong, K.C., *et al.*: 'Collective human mobility pattern from taxi trips in urban area', *Plos One*, 2012, **7**, (4), p. e34487
[7] Ifsttar, M.E.M., Ifsttar, E.C., Ifsttar, J.B., *et al.*: 'Understanding passenger patterns in public transit through smart card and socioeconomic data'. ACM SIGKDD Workshop on Urban Computing, New York, NY, USA, 2014
[8] Ma, X., Liu, C., Wen, H., *et al.*: 'Understanding commuting patterns using transit smart card data', *J. Transp. Geogr.*, 2017, **58**, pp. 135–145
[9] Zhang, F., Zhao, J., Tian, C., *et al.*: 'Spatiotemporal segmentation of metro trips using smart card data', *IEEE Trans. Veh. Technol.*, 2016, **65**, (3), pp. 1137–1149
[10] Mohamed, K., Côme, E., Oukhellou, L., *et al.*: 'Clustering smart card data for urban mobility analysis', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (3), pp. 712–728
[11] Long, Y., Shen, Z.: 'Profiling underprivileged residents with mid-term public transit smartcard data of Beijing', in '*Geospatial analysis to support urban planning in Beijing*' (Springer, New York, NY, USA, 2015), pp. 169–192
[12] Bagrow, J.P., Lin, Y.R.: 'Mesoscopic structure and social aspects of human mobility', *Plos One*, 2012, **7**, (5), p. e37676
[13] Noulas, A., Scellato, S., Lambiotte, R., *et al.*: 'A tale of many cities: universal patterns in human urban mobility', *Plos One*, 2012, **7**, (5), p. e37027
[14] Uppoor, S., Trullols Cruces, O., Fiore, M., *et al.*: 'Generation and analysis of a large-scale urban vehicular mobility dataset', *IEEE Trans. Mob. Comput.*, 2014, **13**, (5), pp. 1–1
[15] Veloso, M., Phithakkitnukoon, S., Bento, C.: 'Sensing urban mobility with taxi flow'. Proc. of the 3rd ACM SIGSPATIAL Int. Workshop on Location-Based Social Networks, Chicago, IL, USA, 2011
[16] Long, Y., Sun, L., Sui, T.: 'A urban research literature review based on public transit smart card data (in Chinese)', *Urban Planning Forum*, 2015, **3**, pp. 70–77
[17] Pelletier, M.P., Trépanier, M., Morency, C.: 'Smart card data use in public transit: a literature review', *Transp. Res. C, Emerg. Technol.*, 2011, **19**, (4), pp. 557–568
[18] Lathia, N., Smith, C., Froehlich, J., *et al.*: 'Individuals among commuters: building personalised transport information services from fare collection systems', *Pervasive Mob. Comput.*, 2013, **9**, (5), pp. 643–664
[19] Briand, A.S., Côme, E., Trépanier, M., *et al.*: 'Analyzing year-to-year changes in public transport passenger behaviour using smart card data', *Transp. Res. C, Emerg. Technol.*, 2017, **79**, pp. 274–289
[20] DBSCAN: 'Density-based spatial clustering of applications with noise'. Proc. of Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, USA, 1996
[21] Ankerst, M.: 'Optics: ordering points to identify the clustering structure', *Stanford Res. Inst Memo Stanford Univ.*, 1999, **28**, (2), pp. 49–60
[22] Kieu, L.M., Bhaskar, A., Chung, E.: 'A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data', *Transp. Res. C, Emerg. Technol.*, 2015, **58**, pp. 193–207
[23] Zheng, B., Zheng, K., Sharaf, M.A., *et al.*: 'Efficient retrieval of top K most similar users from travel smart card data'. 2014 IEEE 15th Int. Conf. on Mobile Data Management, Brisbane, Australia, 2014, vol. 1, pp. 259–268
[24] He, L., Agard, B., Trépanier, M.: 'A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method', *Transportmetrica A: Transp. Sci.*, 2018, pp. 1–20
[25] Rodriguez, A., Laio, A.: 'Clustering by fast search and find of density peaks', *Science*, 2014, **344**, (6191), pp. 1492–1496
[26] Bureau, C.S.S.: 'Statistical yearbook of China 2010', Bureau, China State Statistical, 2010
[27] Bureau, C.S.S.: 'Statistical yearbook of China 2014', Bureau, China State Statistical, 2014


© The Institution of Engineering and Technology 2019

17519578, 2019, 12, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0212 by Beihang University (Buaa), Wiley Online Library on [19/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License