# Perspectives on Stability and Mobility of Passenger's Travel Behavior through Smart Card Data

Zhiyong Cui
Peking University
Beijing, China
1201210546@pku.edu.cn

Ying Long
Beijing Institute of City Planning
Beijing, China
longying1980@gmail.com

## ABSTRACT

Existing studies have extensively used temporal-spatial data to mining the mobility patterns of different kinds of travelers. Smart Card Data (SCD) collected by the Automated Fare Collection (AFC) systems can reflect a general view of the mobility pattern of the whole bus and metro riders in urban area. Since the mobility and stability are temporally and spatially dynamic and therefore difficult to measure, few work focuses on the transition of their travel pattern between a long time interval. In this paper, an overview of the relation between stability and regularity of public transit riders based on SCD of Beijing is presented first. To analyze the temporal travel pattern of urban residents, travelers are classified into two categories, *extreme* and *non-extreme* travelers. We have two lines for profiling all cardholders, rule based approach for extreme and improved density-based clustering method for non-extreme. Similar clusters are aggregated according their features of regularity and occasionality. By combining transition matrix of passenger's temporal travel pattern and socioeconomic data of Beijing in the year of 2010 and 2014, several analyses about resident's temporal mobility and stability are presented to shed lights on the interdependence between stability and mobility in the time dimension. The results indicate that passengers' regularity is hard to predict, extreme travel patterns are more vulnerable and overall non-extreme travel patterns nearly stay the same.

## Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Clustering; H.2.8 [**Database Management**]: Database Applications—*Data mining*

## Keywords

Smart Card Data, Mobility, Stability, SS-OPTICS Clustering

## 1. INTRODUCTION

The continuum of human spatial immobility-mobility at varying geographic and temporal scales poses fascinating topics and challenges for researchers to make right decisions on urban development. Stability and mobility are relative and linked, since mobility reflects movement in short-term temporal or small spatial scales, while stability refer to long-term. Geographically, people move over scales ranging from a few meters to hundreds of kilometers; 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. However, due to the lack of data, research work on mobility and stability is still seldom carried out. Thus, figuring out the puzzle of the relation between stability and mobility will be very meaningful and can help uncover different aspects of public transit, social and urban dynamics.

The temporal and spatial dynamic mobility pattern of residents has been concerned about for a long time by researchers in the fields of transportation engineering[11], computer science [20], urban planning [9], or even socioeconomics [4]. Along with the development of computer science and geographic information (GIS), many new technologies and new types of data can be utilized to measure people's mobility pattern in large-scale regions, such as Call Detail Records of mobile phone [7], taxicabs' GPS information [14], or even outdoor Wi-Fi signal data. When comes to a 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 [5].

AFC systems based on contactless smart cards are available for both city buses and metros to record the details of transaction information when passengers 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 utilize 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 passenger's transit network, reflect his social and economic characteristics, and even forecast his travel pattern. From a city perspective, SCD acting as an transportation probe can help estimate transportation conditions and provide new materials for urban planning policy.


UrbComp'15, August 10, 2015, Sydney, Australia.

In this paper, we utilize the temporal information of SCD to mine the relationship between passenger's mobility and stability in different time and frequency scales. To better understand the passenger behavior in public transportation, we introduce other socioeconomic data into our analysis. Our contributions can be described as follows:

- We take passenger's regularity into account to the analyze relation between regularity and stability.
- We profile passengers with a rule-based classification approach for extreme travelers and an improved density-based clustering method for non-extreme travelers.
- We analyze the mobility and stability of extreme and non-extreme travelers in different group granularities by combining socioeconomic data.

The organization of this paper is as follows: Related work is briefly discussed in section 2. In section 3, we discuss the relation between regularity and stability and profile the passengers. Our analysis about 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

Hanson [4] is among the first researchers to focus on stability and show analyzing individuals' stability requires also analyzing their mobility. Through an empirical example centered on the relationship between entrepreneurship and place, he propose explicitly considering locational stability requires examining stability and mobility in tandem, since spatiotemporal dynamics involved. Based on this idea, James et al. [2] concentrate 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 [11], [12], [18], [19] choose a perspective on large-scale mobility about urban human beings, vehicles or taxis.

To measure residents' stability and mobility in urban area, SCD in public transit is one of the most widely used data. According to Long et al.[10], SCD related research topics can be classified as: 1) data processing and data complementation, like back-calculation of origin and destination and recognition of trip purpose; 2) supporting and management of public transit systems; 3) place-based urban spatial structure and 4) person-based analysis on social network and special group of people. Pelletier et al. [13] also give a literature review of SCD use in public transit and present three levels of management and usage of SCD: strategic (long-term planning), tactical (service adjustments and network development), and operational (ridership statistics and performance indicators). Zheng et al. [20] show us several typical applications based on SCD, like building more accurate route planners. While, Long et al.[9] seek to understand extreme public transit riders in Beijing using both traditional household surveys and SCD. In their work, public transit riders are classified into four groups of different types of extreme transit behaviors to identify the spatiotemporal patterns of these four extreme transit behaviors. Further, Neal et al. [8] discuss personalizing transport information services based on SCD. Among their contributions, the authors use clustering to prove that the usage of public transportation can vary considerably between individuals. Each passenger's trips are aggregated into a weekday profile describing his temporal habits and hierarchical agglomerative clustering is introduced to discover groups of passengers characterizing different travel habits. Contrary to this approach, our weekly profile, presented in Section 3, consisting of hour-grained grid can show more details.

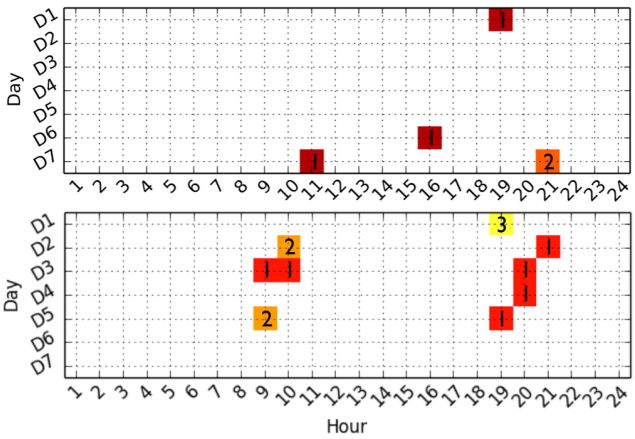


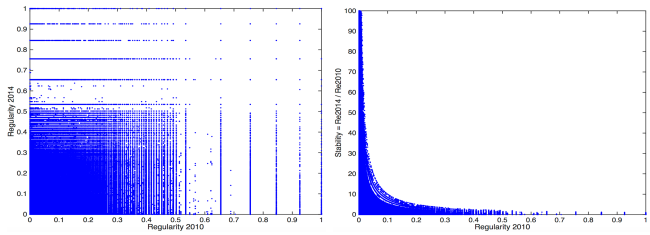

**Figure 1: Weekly profiles of two passengers' transaction time. The transaction time (colored squares) reflects their different travel pattern. Numbers in squares represent times of transaction in that hour. D1-D5: weekdays, D6: Saturday, D7: Sunday.**

As we investigated, many methods and algorithms are adopted to process and analyze SCD. To clustering the temporal information, Mahrsi et al. [5] construct temporal passenger profiles based on boarding information and apply a generative model-based clustering approach to discover clusters of passengers. They also assign passengers based on their boarding information to "residential" areas, which they established through a clustering of socioeconomic data of the Rennes, France, to inspect how socioeconomic characteristics are distributed over the passenger temporal clusters. A density-based clustering method, DBSCAN [3], which is very similar to OPTICS [1] is used by [11]. The authors identify 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. Comparing to approaches reported 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 utilize SCD to measure the interdependence between stability and mobility in the time dimension.

## 3. PROFILING PASSENGERS

### 3.1 Dataset Description

The Smart Card Data (SCD) collected and issued by Beijing Transit Incorporated contains transit riders' records for both the bus and metro systems. There were two types of Automatic Fare Collection (AFC) system on Beijing buses:




(a) Plot of RE10 - RE14 (b) Plot of RE10 - Stability

**Figure 2: Relationship between regularity and stability**

flat fares and distance-based fares, until the beginning of 2015, since when all bus lines became distance-based fare system. 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 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 1) smart card ID, 2) boarding or alighting time, and 3) station ID of boarding or alighting line. As the time span 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 (7×24) 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. Figure 1 illustrates weekly profiles of passengers' transaction time.

## 3.2 Regularity and Stability

In this section, we aim to figure out the relation between passengers' regularity and stability in their daily travel. The large amount of SCD in 2010 and 2014 can help us understand each passenger's weekly travel regularity. We take three aspects of weekly regularity into consideration:

- Travel frequency of the week, $W = \frac{d}{7} \in [\frac{1}{7}, 1]$. Here, $d$ is the number of days when passengers travelled by public transit.
- Travel frequency of every day. We count the number of trips in each day of the week, $D = \{D_i | i = 1, ..., 7\}$. The standard deviation of D is calculated as $D_{sd}$.
- Temporal differences between daily trips. We acquire the temporal differences of $n$ daily trips in one week, $DIST = \{Dist_i | i = 1, ..., \frac{n*(n-1)}{2}\}$, by using the distance calculating method presented in Section 3.4.1. $DIST_{sd}$ is the standard deviation of $DIST$.

Then, since $D_{sd}$ and $DIST_{sd}$ is negative correlated with regularity, we defined passenger's regularity ($RE$) as:

$$RE = W \times e^{-D_{sd}} \times e^{-DIST_{sd}}, RE \in (0, 1] \quad (1)$$

Here, we use the exponential function ($e^{-(x)}$) to normalize the $RE$, ranging from 0 to 1. We also acquire each passenger's stability ($Sta$), which subject to the variance between each passenger's regularities in 2010 and 2014 ($RE10$ and $RE14$), $Sta = RE14/RE10$. Figure 2(a) shows the relation between $RE10$ and $RE14$, and the correlation coefficient is 0.0485. Figure 2 (b) shows the relation between $RE10$ and $Sta$, and the correlation coefficient is -0.00059. This two coefficients are both less than 0.1, which means the regularities of passengers between 4 years are nearly irrelevant. We may assert that the regularity between long-time intervals cannot be predicted.

## 3.3 Identifying Extreme Travelers

Before analyzing the transit behaviors of the passengers in Beijing, we separate the whole passengers into two groups: *extreme* travelers and *non-extreme* travelers. We define and identify the extreme travelers according to a survey in 2010 [9] as well as researchers' own experiences of living in Beijing. Four types of extreme travelers are defined based on their behaviors in weekdays, by setting several thresholds and combining empirical knowledge of Beijing as depicted in Table 1. For example, since most people's working hours start on 8:30 or 9:00 am in Beijing, public transit boarding time before 6:00 am would be considered as an unusually early situation.

**Table 1: Definitions of extreme travelers**

| Type | Defination |
|---|---|
| Early Birds (EBs) | First trip < 6AM, more than two days in five weekdays (60% of weekdays) |
| Night Owls (NOs) | Last trip > 10PM, 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) |

## 3.4 Clustering Non-extreme Travelers

As extreme travelers only account for a small proportion (less than 5%), we cluster the non-extreme travelers to character their travel pattern. This process is consisted of three stages: 1) defining the distance (similarity) between different SCD records; 2) clustering samples of SCD with a simplified-smoothed OPTICS algorithm proposed by us; and 3) classifying the whole SCD records with a Kmeans-like algorithm according to results of the clustering stage.

### *3.4.1 Defining Distance between Smart Card Records*

We count the transaction time of SCD in each hour of the week to form a vector consisted of 168 (24 hours × 7 days) variables, $V = [v_0, \cdots, v_{167}] \in N$. There are some classic distance-measurement methods to measure the similarity of different records, like Euclidean distance, Manhattan distance, and cosine distance. As we tested, given two vectors,

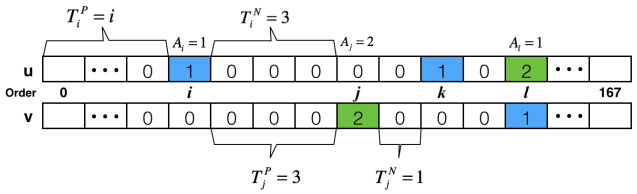



**Figure 3: Distance between two vectors, $u$ and $v$, is the sum of the time interval ($T_i = min\{T_i^P, T_i^N\}$) and absolute difference between $u_i$ and $v_i$ ($A_i$) in each position $i$, where $u_i \neq v_i$.**

Euclidean distance and Manhattan distance only compute the sum of differences between components in the same position of two vectors. But they will not consider the influence of components' positions which reflect vectors' temporal attribute. When computing cosine distance, since the vectors are mostly sparse, the product of two components, one is zero and the other is not, in the same position of two vectors will be zero. This will miss out many useful information of the two vectors. Thus, those classic distances formulas are not capable of measuring the time interval between smart card transactions.

To solve the above problems, we define a method to compute the distance between two vectors as **Transaction Distance** ($D_{tran}$), considering both vector's difference and temporal attribute. Since non-extreme passengers' vectors are mostly sparse vectors. We define the distance between the two vectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, by computing the sum of $i$th component distance ($D_i$) between $u_i$ and $v_i$. The component distance ($D_i$) consists of two parts, the time interval ($T_i$) and the absolute difference of the two components' value ($A_i = |u_i - v_i|$). As for the time interval $T_i$, if one of $u_i$ and $v_i$ equals to 0, $T_i$ equals the smaller value of the previous and the next time intervals between non-zero components in different vectors, namely $T_i = \min\{T_i^P, T_i^N\}$. If $u_i$ and $v_i$ both do not equal to 0, $T_i = 0$. Then, the Transaction Distance between vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ can be represented as:

$$D_{tran} = \sum_{i=0}^{167} \min\{T_i^P, T_i^N\} + k * |u_i - v_i|, \quad s.t. \quad u_i \neq v_i \quad (2)$$

Here, $k$, ranging from 0 to 3, is a parameter to balance the weights of $T$ and $A$, as we tested. Figure 3 shows an example of computing the transaction distance. $T_j = min\{T_j^P, T_j^N\}$ equals 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 $u_i$, its $T_i^P$ equals $\min\{i, 167 - i\}$.

### 3.4.2 Clustering Samples of SCD Records

We then cluster the vectors to identify the travel pattern of public transit riders in Beijing based on the distances of smart card transaction records. Although K-Means algorithms or other centroid models are very efficient to cluster the travelers pattern, it is hard to nominated the number of clusters (k) before running of the algorithm, without prior knowledge. A new fast-searching and density-based clustering algorithm [15] can only identify 4 or 5 obvious clusters as we tested. Thus, we propose an improved density-based

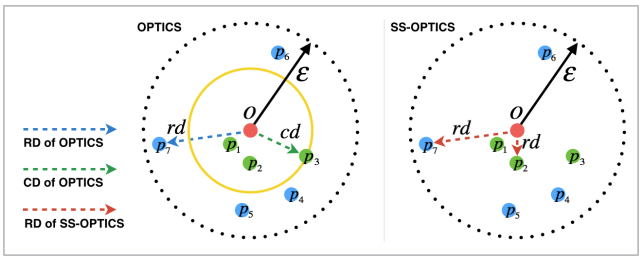



**Figure 4: cd of OPTICS and rd of both OPTICS and SS-OPTICS. *MinPts* = 4**

clustering algorithm based on OPTICS [1], which is suitable for our data with distances acquired. We named it as Simplified-Smoothed-OPTICS (SS-OPTICS).

*1) Simplify*: The original OPTICS algorithm has two key concepts, *cord-distance* and *reachability-distance*.

**Definition-1, *Core-distance (cd)***: Let $p$ be an object from a dataset D, let $\varepsilon$ be a distance value, let $N_\varepsilon(p)$ be the set $\{x \in D | dist(p, x) \leq \varepsilon\}$, let *MinPts* be a natural number and let *MinPts-distance(p)* be the distance from $p$ to its *MinPts* neighbor. Then, the *core-distance* of $p$ is defined as *core-distance*$_{\varepsilon, MinPts}(p)$ =

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

**Algorithm 1:** Getting Ordered Points by OPTICS

```
Data: D (Unprocessed Dataset), ε
Result: OrderedPoints
initialization;
while D ≠ Null do
    Point = D.pop();
    OrderedPoints.append(Point);
    P_neighbors = point.neighbor(ε);
    if P_neighbors ≠ 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(ε);
            if S_neighbors ≠ Null then
                OrderSeeds.updateRD(Seed, S_neighbors)
```

**Definition-2, *Reachability-distance (rd)***: Let $p$, $o \in$ D, let $N_\varepsilon(o)$ be the $\varepsilon$ -neighborhood of $o$, let *MinPts* be a natural number. Then, the *reachability-distance* of $p$ with respect to $o$ is defined as *reachability-distance*$_{\varepsilon, MinPts}(p, o)$ =

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

Here, $\varepsilon$ and *MinPts* are two input parameters of the original OPTICS algorithm. According to OPTICS's definitions, the

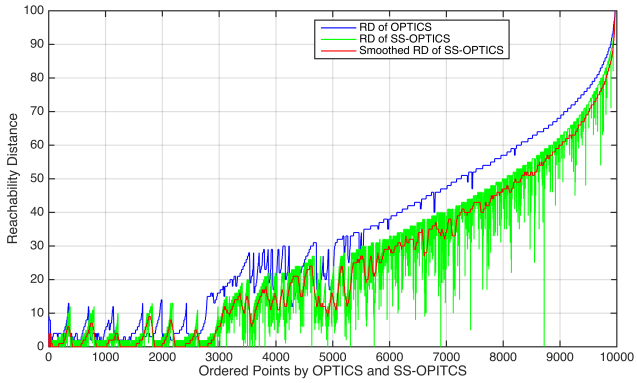



**Figure 5: RD curves of OPTICS and SS-OPTICS, $\varepsilon$ = 100 and S = 41**

green points covered by the yellow circle in the Figure 4 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 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: 1) reachability distance represents density and 2) 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 Figure 4 maybe disordered due to their same $rd$s. Thus, we design an improved OPTICS algorithm, mainly shown in Algorithm 1, by abandoning the concept of *core-distance* and define a new concept of *reachability-distance (RD)* as follows.

**Definition-3, *New Reachability-distance (RD)***: Let $p$, $o$ $\in$ D, let $N_\varepsilon(o)$ be the $\varepsilon$ -neighborhood of $o$. The *reachability-distance* of $p$ with respect to $o$ is defined as *reachability-distance*$_\varepsilon$*(p,o)* =

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

*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 Figure 5, the blue and green lines are the $rd$ curves of OPTICS and SS-OPTICS, respectively. We notice that, although the value of SS-OPTICS's RD is less than OPTICS's, their curves are extremely similar.

The red line is the smoothed RD of SS-OPTICS, $RD'$, in Figure 5. We smooth the RD curve with two aims: 1) easily identifying the the valley-shaped clusters and 2) controlling the size of a cluster. We use mean filter to smooth the RD curve to achieve our goals with only one parameter, window size (S). Each value of the smoothed RD curve, $RD'_i$, is the mean of RD value of points within the window:

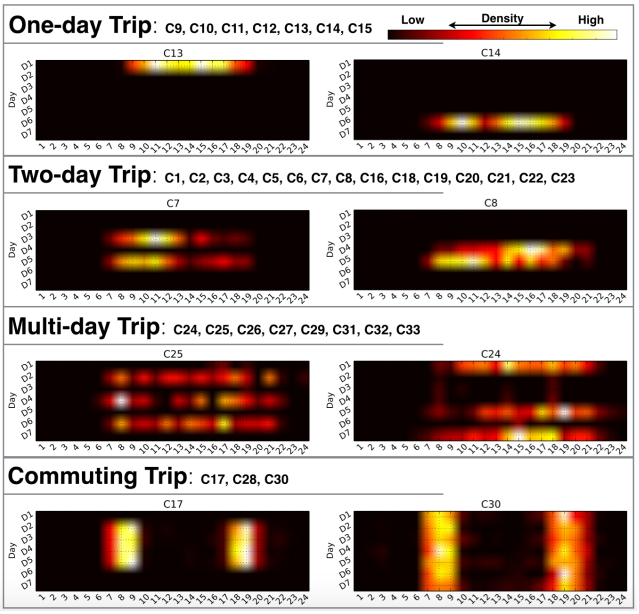



**Figure 6: Four obvious categories of the heatmap of the 33 clusters. D1-D5: weekdays, D6-D7: weekends**

$$RD'_i = (\sum_{j=i-n}^{j=i+n} RD_j)/S, \quad s.t. \quad n = \frac{S-1}{2} \tag{3}$$

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 $\frac{S-1}{2}$. After the mean filtering, the valley (cluster) of the RD curve, whose number of the points in this cluster is less than $\frac{S-1}{2}$, will nearly be filled up. Thus, the cluster size is controlled to be larger than $\frac{S-1}{2}$.

As we tested, the average sizes of clusters generated by SS-OPTICS is 10% larger than that of OPTICS and the average cohesion of clusters generated by SS-OPTICS is around 3% smaller than that of OPTICS. Further, the results of clustering by the two methods are nearly the same, if the input points are distributed in normal shapes, like square, circle or Gaussian. And both SS-OPTICS and OPTICS are not sensitive to the value of input parameter with time complexity of $O(n^2)$. But SS-OPTICS only needs one parameter ($\varepsilon$, setted as 100 in our experiment), while OPTICS needs two ($\varepsilon$ and *MinPts*). Meanwhile, SS-OPTICS is more easier to control the cluster size by defining the value of $S$. Finally, we iteratively cluster several random samples of SCD, containing 20000 entries in each sample, and identify 33 clusters for the next stage to classify the whole dataset.

According to the transaction time distribution of the 33 clusters, they can obviously be classified into 4 big categories as Figure 6 shows. The 4 categories can be described as: *one-day trips*, *two-days 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). The commuting trips (clusters 17, 28 and 30) are mainly charactered with regular transaction time distributed in weekdays.

### 3.4.3 *Classifying the Whole SCDs*

The 33 clusters acquired by SS-OPTICS are described as $C = [C_1, ..., C_{33}]$. Each $C_i$ in $C$ is a one-dimensional vector containing 168 components. Each component $(c_j)$ of $C_i$ is the incidence rate of passenger's smart card transaction in the $(j\%24)$th hour of the $(j - j\%24/7)$th 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. Thus, we can classify all the SCDs based on the clusters' feature, $C = [c_{ij}]_{34\times168}$.

According to the data we already known: 1) cluster number $k$ and 2) feature of each cluster $C_i$, it is very suitable for us to utilize Kmeans-like algorithm to classify the whole dataset, since the nodus of Kmeans is to fix k and the centroid of each cluster. For each SCD vector, $V = [v_0, \cdots, v_{167}]_{v_i \in N}$, it belongs to Cluster $C_i$:

$$i = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{168} v_j \times c_{ij} \tag{4}$$

Then, we update the cluster $C_i$, $c_{ij}=$

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

Here, $n$ is the total number of transactions in $C_i$.

## 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 many work [11], [5], [13] mentioned, utilizing 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. But we still need to consider the influence of many other factors, like age distribution, social scale, per capita income, type of job, city size and so on, to analyze and make correct decisions. Since the datasets of 2010 and 2014 are selected according to same smart card IDs, the mobility and stability of fixed passengers can be reflected by their changes of travel pattern between the two years. In this section, we analyze passenger's mobility and stability pattern based on temporal information combining some background socioeconomic factors listed in Table 3.

### 4.1 Relation between Mobility and Stability

Different temporal scales of data reflect facts from different perspectives. Weekly profiles showing short-term mobility depict people's living circles, while transitions of mobility patterns may imply the unchangeable of lifestyle and social status between several years. A case study of variability of temporal patterns in Singapore [21] shows variability of mobility patterns can be observed at individual and spatial aggregated scale, but the overall urban movements remains almost the same. Jiang et al. [6] cluster individuals' daily activity patterns according to their usage of space and time within one year, and show that daily routines can be highly predictable at a group scale. But the analysis of relation between regularity and stability in Section 3.2 shows it is hard to predict passenger's regularity. Thus, the predictability of passenger's regularity is likely to be influenced by the time interval of prediction.

Meanwhile, our long-term analysis about mobility and stability can be mutual corroborated to some extent by the above two works focusing on short-term analysis [21], [6]. As we can see in the fine-grained and coarse-grained comparisons of non-extreme passengers' transit profiles between 2010 and 2014 in next sections, long-term dynamics of extreme travelers give us a snapshot of urban dynamics. Along with the increase of population and urban size in Beijing, inhabitant's travel pattern changes a lot. But people's life styles are more or less at a standstill. They still live in such a big city, they still have to use public transit systems, and they still have no more choice but to ride buses or metros. The passenger's performance shows how difficult to ascend a higher stratum of the society in leaps and bounds in four years. The relation between mobility and stability also makes us have a better understanding of that people and society advance by steps not by leaps.

**Table 2: Transition Matrix of Extreme Travelers**

| 2010 \ 2014 | EB | NO | TI | RI | NE | SUM |
|---|---|---|---|---|---|---|
| EB | 1286 | 206 | 535 | 82 | 7605 | 9714 |
| NO | 299 | 2550 | 2200 | 153 | 30006 | 35208 |
| TI | 376 | 996 | 9488 | 182 | 48406 | 59448 |
| RI | 93 | 198 | 677 | 275 | 7351 | 8594 |
| NE | 8780 | 26357 | 82630 | 3977 | 1646118 | 1767862 |
| SUM | 10834 | 30307 | 95530 | 4669 | 1739486 | 1880826 |

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

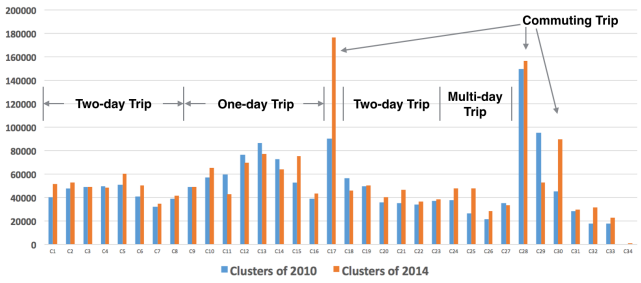



**Figure 7: Four obvious categories of the heatmap of the 34 clusters**

### 4.2 Extreme Travelers Analysis

According to the classification criteria proposed in Table 1, we get the transition matrix of the 4 types of extreme travelers (EB, NO, TI, RI) from 2010 to 2014 in Table 2. The amounts of extreme travelers in 2010 (141340) and 2014

**Table 3: Social Economics Factors in Beijing**

| *Year* | *Population* | *Population Density* | *Private Vehicles* | *Bus Volume* | *Metro Volume* |
|---|---|---|---|---|---|
| 2010 | 17.55 *mil.* | 1224 $persons/km^2$ | 2.97 *mil.* | 5.165 *bil.* | 1.423 *bil.* |
| 2014 | 21.15 *mil.* | 1498 $persons/km^2$ | 4.25 *mil.* | 4.843 *bil.* | 3.205 *bil.* |

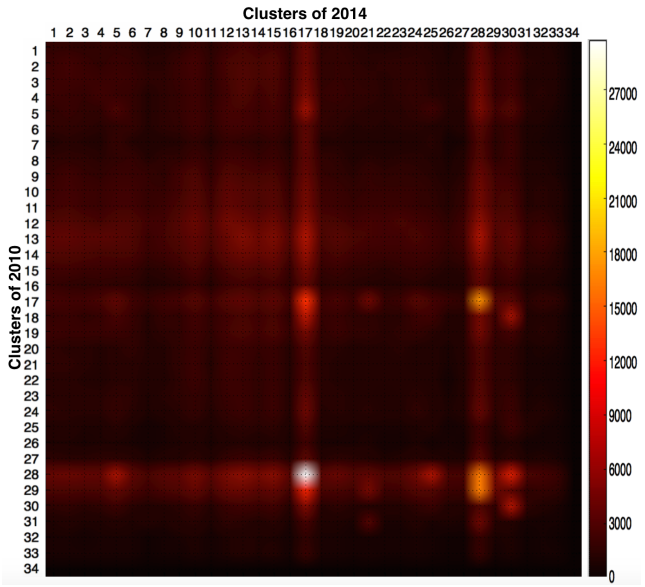



**Figure 8: Heatmap of the 34 clusters' transition matrix**

(112964) are both very small comparing to that of non-extreme travelers. In addition, 84% of the extreme travelers in 2010 converted into non-extreme travelers in 2014, which means the stability of extreme travelers' live pattern cannot last for a long time.

But among the 4 extreme type's transition, we still find that the amounts of EB, NO and TI in 2010 converted to themselves in 2014 (1286, 2250 and 9488) occupy the highest rate. That is to say, extreme pattern is more likely to keep the original status other than to convert into other extreme patterns. It also meets the findings of our previous work [9] 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. The phenomenon of the transition rate of TI to TI (86%) greatly exceeds that of TI to EB, NO or RI demonstrates that people in the group of TI may have greater difference of work patterns compared to others.

## 4.3 Non-Extreme Travelers Analysis

As Figure 6 illustrates, the heatmaps of non-extreme clusters can be classified into the 4 categories. Heatmap's feature in each category is so distinctive that it seems like that these clusters are classified by some thresholds or a decision tree, which reflects the accuracy of our clustering method. After clustering the sample data and classifying the whole dataset, we get the amounts of cards in each of the 34 clusters in 2010 and 2014, demonstrated by Figure 7. The amounts of transaction time in *one-day*, *two-day* and *multi-day* trips do not vary much between 2010 and 2014. The amount of of one-day trips, around 60000, is a little more than that of two-day trips, around 50000. The amount of multi-day trips (about 30000 in each cluster) is the least. This phenomenon may be explained as: the more disperse the passenger's trips are in one week, the smaller the amount of this kind of passengers is. Except for the workers commuting by public transit, only a small portion of people travel a lot with their travel time irregularly distributed in the week. These data also shows more people in Beijing choose to use public transit occasionally, mainly in one day or two days. It is maybe related to the huge urban size which leads residents will not use public transit only if they go far.

However, almost all the towering bars in Figure 7 belong to commuting trips, which depicts the public transit commuters who take a home-to-work trip every weekday morning and go back to home in the evening. Under scrutiny, the amount of passengers belonging to commuting trips nearly doubled from 2010 to 2014. To explain this evident increase, two reasons should be considered. One is that public transit became more convenience from 2010 to 2014, since Beijing metro company constructed 8 more lines into 15 lines in total and the total metro length increased rapidly from 228 km to 465 km during this 4 years. The other reason is the ground transportation in Beijing became more congested, since the total number of private vehicles in Beijing increased from 2.9 to 4.3 million.

### 4.3.1 Fine-grained Analysis

By acquiring the amounts of passengers of the 34 clusters in 2010 and 2014, we calculate the transition (mobility) matrix of these clusters demonstrated by a heatmap shown in Figure 8. In this heatmap, the brighter the grid is, the more passengers belong to this grid. We can easily catch sight of bright parts (red, orange, yellow and white parts) and find these parts mainly distributed in cluster 17, cluster 28 and cluster 30 of both 2010 and 2014, which belong to the commuting trips category.

Especially for the yellow and white grids ($C_{17\to17}$, $C_{17\to28}$, $C_{28\to17}$ and $C_{28\to28}$), their amounts are several times as large as the amounts of other grids. This reflects the stability of people belonging to commuting trips category, who tend to remain the same status. The four grids' weekly profiles are demonstrated by heatmaps in Figure 9. Although their morning-evening rush hours have a deviation of one hour, the stability can be reflected by the similar distribution of transaction time and the same time intervals between morning and evening rush. Their temporal profiles also tell us most commuting trips of passengers in Beijing are distributed mainly from Tuesday to Friday. It is interesting to see why commuting passengers use public transit

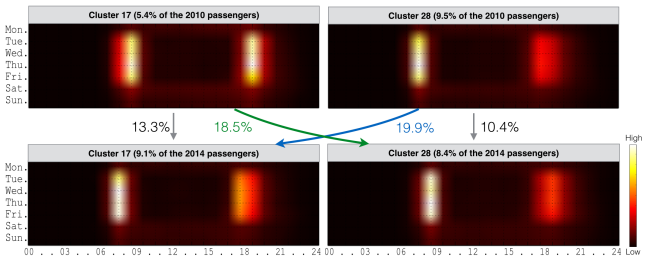



**Figure 9: Heatmaps of the mutual transition between cluster 17 and cluster 28 in both 2010 and 2014**

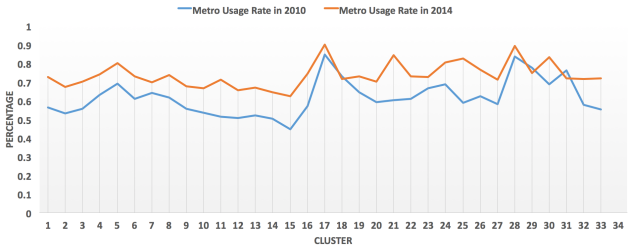



**Figure 10: Ratios of the passengers riding in a metro at least once a week**

on weekdays except Monday in the future. A possible explanation may be the Monday Morning Syndrome (MMS), which means some people feel even more tired out than on Friday after relaxation over the weekend.

There are also some red and orange 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 job or work location, earning enough money to buy a car, or taking metro to work instead of driving. Figure 10 shows the ratios of passengers who rode in a metro at least once a week in each cluster. The ratio in 2014 is apparently higher than that of 2010. Further, the ratios of commuting clusters (17, 28, and 30) reach local peaks in both lines of 2010 and 2014 in the figure. This means commuting passengers may be the most stable group who are most willingly to transit by metro.

### 4.3.2 *Coarse-grained Analysis*

The transition matrix of the 4 types of non-extreme travelers is also counted according to above data, in Table 4, and give us a new perspective to analyze passenger's mobility and stability. Here, people belonging to $O$ and transformed into $C$ is represented by $T_{O\to C}$ as a component of transition matrix. However, only with smart card data, we cannot prove our conjectures. To better understand the mobility and stability of passengers, we combine some socioeconomic statistics data of Beijing in both 2010 and 2014 [16], shown in Table 3. From 2010 to 2014, Beijing increased 3.6 million

**Table 4: Transition Matrix of Non-extreme Travelers**

| 2010 \ 2014 | O | T | M | C | SUM |
|---|---|---|---|---|---|
| O | 119270 | 193290 | 84311 | 31864 | 428735 |
| T | 164817 | 298667 | 142436 | 48043 | 653963 |
| M | 64449 | 142399 | 76769 | 20038 | 303655 |
| C | 36642 | 79266 | 47757 | 11812 | 175477 |
| SUM | 385178 | 713622 | 351273 | 111757 | 1561830 |

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

people and the population density in urban area rose from 1224 to 1498 persons per square kilometer. Along with the growth of population, the total number of private vehicles in Beijing expanded to 4.25 million from 2.97 million. All these factors tell us a fact that Beijing became more crowed in urban area and more vehicles led more congested ground transportation in 2014. 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 population and vehicles increased a lot in Beijing, data reveals the stability of travel pattern of public transit riders. However, as table 3 indicates, the total volume of Beijing Metro System doubled during the 4 years, while the volume of bus system decreased. This unusual decline can be explained that the government focused on pushing forward the expansion of the Beijing Metro System to mitigate congestion brought by the fast increasing population and vehicles. But the transition matrix uncover that the increased metro lines cannot fix the root cause of this problem. In conclusion, like the findings in [21] and [17], the coarse-grainded analysis shows the mobility of a part of residents may change, but the overall travel patterns of passengers nearly stay the same.

## 5. CONCLUSIONS

Smart card data gives us another opportunity to observe the operation of our cities, where moving is the perpetual normal. In this paper, we analyze the relation between passenger's regularity and stability. For the non-extreme travelers, we cluster them by utilizing SS-OPTICS, proposed by ourselves, to discover their transition patterns between different clusters. By combining socioeconomic data, we present several analyses about resident's temporal mobility and stability. Extreme travelers are most vulnerable that the stability of extreme travelers' life pattern cannot last for a long time. According to clustering outcomes and our analyses, non-extreme travelers show their high mobility by a lot transition between different fine-grained clusters. However, the stability of their travel patterns is also obvious when coarse-grained classification is introduced to our analysis. Along with the increase of population and vehicles in Beijing, although the government constructed more metro lines to mitigate congestion, it cannot solve this problem and the overall public transit riders nearly keep the same travel patterns. But at individual level, a passengers' transit trip is hard to predicted based on short-term travel behavior, like

predicting a passenger's regularity in 2014 based on that in 2010. Further, the prediction will be more difficult for extreme travelers.

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 data can be introduced to our analysis.

## 6. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (No.51408039).

## 7. REFERENCES

[1] M. Ankerst. Optics : Ordering points to identify the clustering structure. *Stanford Research Inst Memo Stanford University*, 28(2):49–60, 1999.

[2] J. P. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *Plos One*, 7(5):e37676, 2012.

[3] DBSCAN. Density-based spatial clustering of applications with noise. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 1996.

[4] S. Hanson. Perspectives on the geographic stability and mobility of people in cities. *Proceedings of the National Academy of Sciences*, 102(43):págs. 15301–15306, 2005.

[5] M. E. M. (Ifsttar, E. C. (Ifsttar, J. B. (Ifsttar, L. O. (Ifsttar, and E. C. (Ifsttar. Understanding passenger patterns in public transit through smart card and socioeconomic data. *Acm Sigkdd Workshop on Urban Computing*, 2014.

[6] S. Jiang, J. Ferreira, and M. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.

[7] C. Kang, S. Sobolevsky, Y. Liu, and C. Ratti. Exploring human movements in singapore: a comparative analysis based on mobile phone and taxicab usages. *Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages - ResearchGate*, 2013.

[8] N. Lathia, C. Smith, J. Froehlich, and L. Capra. Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, 9(5):643–664, 2013.

[9] Y. Long, X. Liu, J. Zhou, and Y. Chai. Early birds, night owls, and tireless/recurring itinerants: An exploratory analysis of extreme transit behaviors in beijing, china. *arXiv preprint arXiv:1502.02056*, 2015.

[10] Y. Long, L. Sun, and T. Sui. A review of urban studies based on transit smart card data (in chinese). *Urban Planning Forum*, 3(10):70–77, 2015.

[11] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1 – 12, 2013.

[12] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *Plos One*, 7(5):: e37027., 2012.

[13] M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557 – 568, 2011.

[14] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò. Collective human mobility pattern from taxi trips in urban area. *Plos One*, 7(4):: e34487., 2012.

[15] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[16] C. Statistical Bureau. Statistical yearbook of china 2010-2014. *Bureau, China. State Statistical*, 2014.

[17] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34):13774–13779, 2013.

[18] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas. Generation and analysis of a large-scale urban vehicular mobility dataset. *IEEE Transactions on Mobile Computing*, 13(5):1–1, 2014.

[19] M. Veloso, S. Phithakkitnukoon, and C. Bento. Sensing urban mobility with taxi flow. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 2011.

[20] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: Concepts, methodologies, and applications. *Acm Transactions on Intelligent Systems and Technology*, 5(3):222–235, 2014.

[21] C. Zhong, E. Manley, S. M. Arisona, M. Batty, and G. Schmitt. Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, 9:125 – 130, 2015.