Characterizing Evolution of Extreme Public Transit Behavior Using Smart Card Data

Zhiyong Cui
Department of Civil and Environmental Engineering
Univesity of Washington
Seattle, WA, US
Email: zhiyongc@uw.edu

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

Ruimin Ke
Department of Civil and Environmental Engineering
Univesity of Washington
Seattle, WA, US
Email: ker27@uw.edu

Yinhai Wang
Department of Civil and Environmental Engineering
Univesity of Washington
Seattle, WA, US
Email: yinhai@uw.edu

***Abstract*—Existing studies have extensively used temporal-spatial data to mine 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. Most existing work focusing on mobility pattern usually ignore a special group of people who travel in abnormal patterns or mechanisms. In this paper, we focus on the evolution extreme transit behaviors of travelers in urban area by using SCD in 2010 and 2014. We have several aspects of descriptive statistics of the SCD with a view to better understanding the dynamic process and evolution of the extreme transit behavior. By combining the SCD's temporal information with the amount of travel behavior, we also propose a concept of Extreme Index (EI) based on the mixture Gaussian model to depict the extreme level of the passengers' travel pattern. According to our analysis, the normal transit behavior of the two years have nearly the same temporal distribution. Although the EI models of the two years have similar distributions, the EI model of 2010 with two peaks is more scattered than that of 2014, which has only one peak. The EI model, which assigns an EI attribute for each SCD, can be applied in further analysis of urban transit or passengers' behavior.**

I. 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 right decisions on urban development and traffic assignment. As a major urban transit method, the public transit has made a great contribution to mitigate congestion and save energy. Under this background, many works are carried out focusing on the analysis of passengers' travel pattern. But there are some extreme transit behaviors we need to pay more attention, since these behaviors always reflect the extreme events and mechanisms in the city.

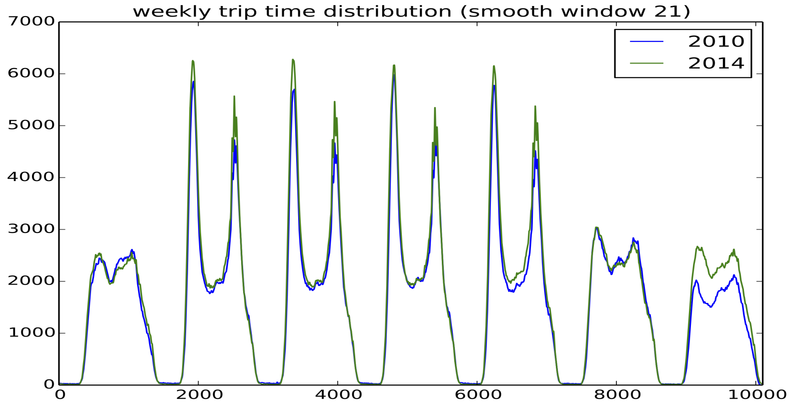
Extreme travelers have received increasing attention in recent years, when we are experiencing the backdrops of the global financial crisis, increased numbers of the unemployed, rise of telecommuters and low-paying jobs relocated to cheaper places inside or outside a region/country. But it needs a more general criteria to evaluate these phenomenon and estimate the evolving of these special travel patterns. In this paper, we focus on the evolution of extreme transit behaviors of travelers in urban area by using SCD in 2010 and 2014. We have several aspects of descriptive statistics of the SCD with a view to better understanding the evolution of the extreme transit behavior. We also proposed a concept of Extreme Index (EI) based on the mixture Gaussian model to depict the extreme level of the passengers' travel pattern.

The paper is organized as follows. In the next section, we describe some previous work related to extreme transit behavior. We then give a brief dataset description in Section III. Several descriptive statistics of the SCD are demonstrated and analyzed in Section IV. Section V mainly characterizes the level of extreme transit behavior and define the Extreme Index (EI). We conclude this paper and have a discussion of future work in Section VI.

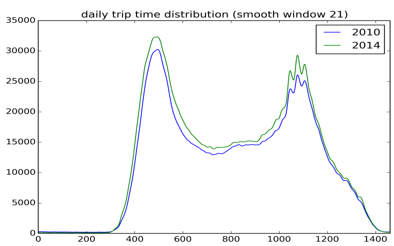
II. Related Work

Many existing works [3] [4] focus on revealing the regularity and predictability of human behavior, which is an opposite view of extreme transit behavior. James et al. [1] 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 [9], [10] choose a perspective on large-scale mobility about urban human beings, vehicles or taxis.

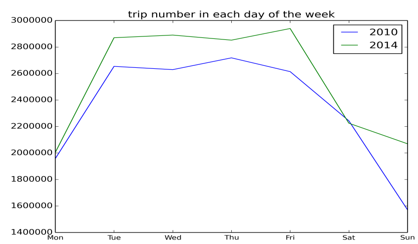
To measure residents' transit behavior in urban area, SCD in public transit is one of the most widely used data. Cui et al. [2] have measured the stability and mobility of passenger's travel behavior through SCD. In their work, public transit riders are clustered into different groups of different types of transit behaviors to identify the extreme transit behavior patterns and estimate the relationship between mobility and stability. Neal et al. [5] discuss personalizing transport information services based on SCD. Zheng et al. [11] show us several typical applications based on SCD, like building more accurate route planners. Further, Legara et al.[6] infer passenger type from commuter eigentravel matrices by focusing on passengers' usual and unusual travel patterns. To the best of our knowledge, Long et al.[7] seek to understand extreme public transit riders in Beijing using both traditional household surveys and SCD and we combine their work to show our findings in this paper.



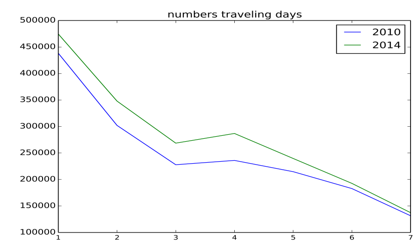
(a) Weekly Trip time distribution of 2010 and 2014. X-axis represents the time lasting for one week with 1 minute interval. Y-axis indicates the number of trips. The curves are smoothed by Hanning function with the window size of 21. Monday in each year is a holiday



(b) Weekly Trip time distribution. X-axis & Y-axis are with the same meanings of Fig.1(a)



(c) Distribution of numbers of trips in each day



(d) Distribution of numbers of smart card holders whose traveling days ranging from 1 to 7 in one week. Y-axis is the number of card holders

Fig. 1. descriptive statistics of the data in 2010 and 2014

III. Dataset Description

To evaluate the evolution of extreme transit, we look into movements of public transit users in Beijing by using smart card data. The SCD contains transit riders' records for both the bus and metro systems. There were two types of AFC system on Beijing buses: flat fares and distance-based fares. In this paper, no matter what kind of fare system, we consider the transaction (paying) time as the time for one ride for simplicity.

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 come from the first complete week in April. The two datasets 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. Moreover, even though someone discards the smart card or returns it back to public transit company, the smart card's ID will not be assigned to others any more 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 we do not have the longitudinal information of public transit stations, we only use the temporal information of passengers' boarding and alighting time for each ride.

IV. Data Statistics & Analysis

To better understand the travel pattern of public transit user, we carry out several aspects of descriptive statistics of the data. Although we do not have the spatial information of the SCD, to characterize the passengers' extreme travel activity, we first introduce the trips' temporal distribution and card holders' trip number distribution into our analysis.

A. *Temporal distribution analysis*

We count smart card's travel activities (trips) with an interval of 1 minute and thus, the whole smart card records can be represented as a vector with 10080 (7*24*60) components. Figure 1(a) shows the weekly trip time distributions of the SCD in 2010 and 2014. The two years' curve are very similar during the whole week. There are 7 obvious curves in the figure which demonstrate the 7 days' trip pattern. Each of the 7 curve has two peaks which represent the morning peak and the evening peak. One special case we need to look at is the first curve for Monday, which is obviously different from the other weekdays. As we checked, the first Mondays of April

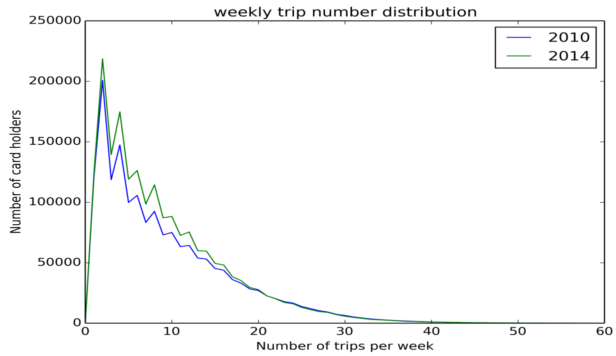


Fig. 2. Trip number distribution with respect to smart card holders during one week

in both 2010 and 2014 are the part of a national holiday of China, the Tomb Sweeping Day. Thus, it is reasonable that the trip amount of Monday is less that that of the other weekdays.

Figure 1(b) shows the distribution of daily travel time during the whole week in each year of 2010 and 2014. The curves have been smoothed by the Hanning function with a window size of 21. The curve nearly starts from 5:00am (the 300th point in the figure) and lasts until the end of the day (the 1400th point). There are two obvious peaks in each curve and the evening peak is scattered with 3 sub-peaks which demonstrates there are mainly 3 kinds of time for workers leaving office. That may be one of the features of a megacity, like Beijng, which has different leaving office time to spread the evening peak over several hours to mitigate the traffic congestion. The distribution of trip numbers in each day of the week is demonstrated in Figure 1(c). The trips of each day, except Saturday, in 2014 are more than that of 2014, which indicates the people holding the smart cards in our dataset increase their public transit travel times after 4 years. Figure 1(d) shows the distribution of numbers of traveling day for each smart card holder. We may expect the curve in this figure would be a monotonically decreasing function. But actually, people who traveling in 4 days a week is more than that of 3 days a week. Considering Monday is a holiday and only 4 days (Tuesday to Friday) are weekdays in this case, the commuters with high possibility to travel in 4 days of the week, accounting for a large proportion of the whole dataset, are one of the major groups of people we need to analysis.

B. Travel times distribution analysis

After we count the travel times of all the card holders, we get a trip number distribution shown in Figure 2. The number of holders who only have 1 public transit trips per week reach the peak of the curves. Then the number of holders decreases as the number of trip increase. One interesting phenomenon in this figure is that both the curves appear to be serrated and the card holder number with odd trip number is mostly smaller than that with the even trip number adjacent. It seems like if people choose to travel by public transit, they mostly would like to travel a round trip other than a single trip.

V. Characterize the Level of Extreme Transit

Our previous work [7] has proposed several definitions of extreme transit behavior. We first define and identify the extreme travelers here. But several kinds of extreme transit behavior cannot completely reflect passengers? extreme travel pattern. Thus, we then propose a more general definition with regard to the extreme pattern based on the SCD analysis above.

A. Four Extreme Transit Behavior

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 I. 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.

According to [7], [8], commuting journeys, which are required by the TIs, can be constructed based on commuters' job and home location. Even though our dataset does not have spatial information about bus/metro station, we can substitute the station's number for the location to identify passenger's job and home location.

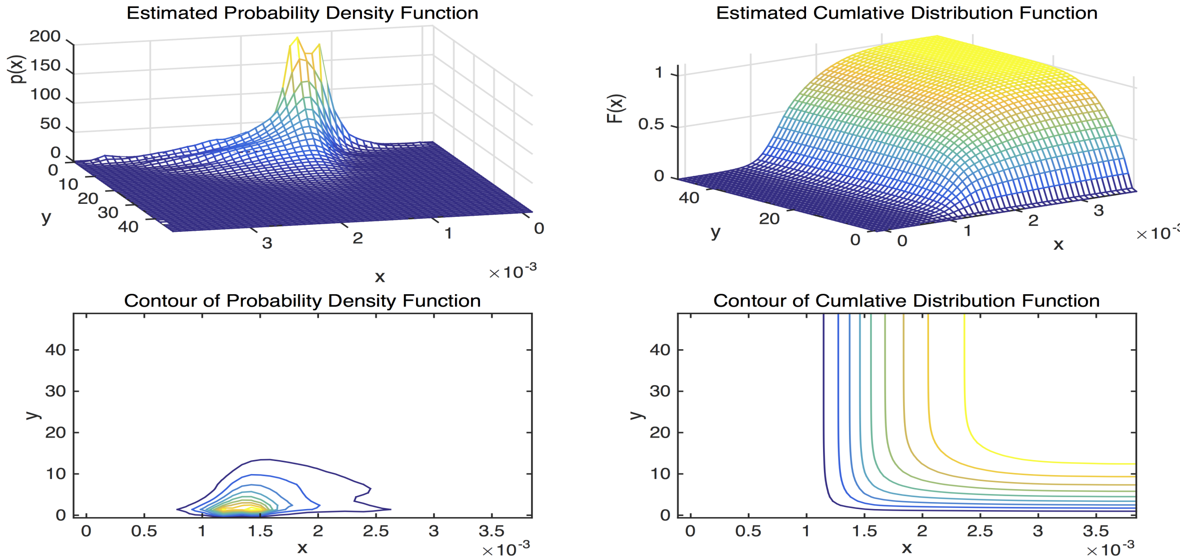
TABLE I
Definitions of extreme travelers

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)	$\geq$ one and a half hours commuting, more than two days in a week
Recurring Itinerants (RIs)	$\geq$ 30 trips in weekdays of a week ($\geq$ 6 trips per day)

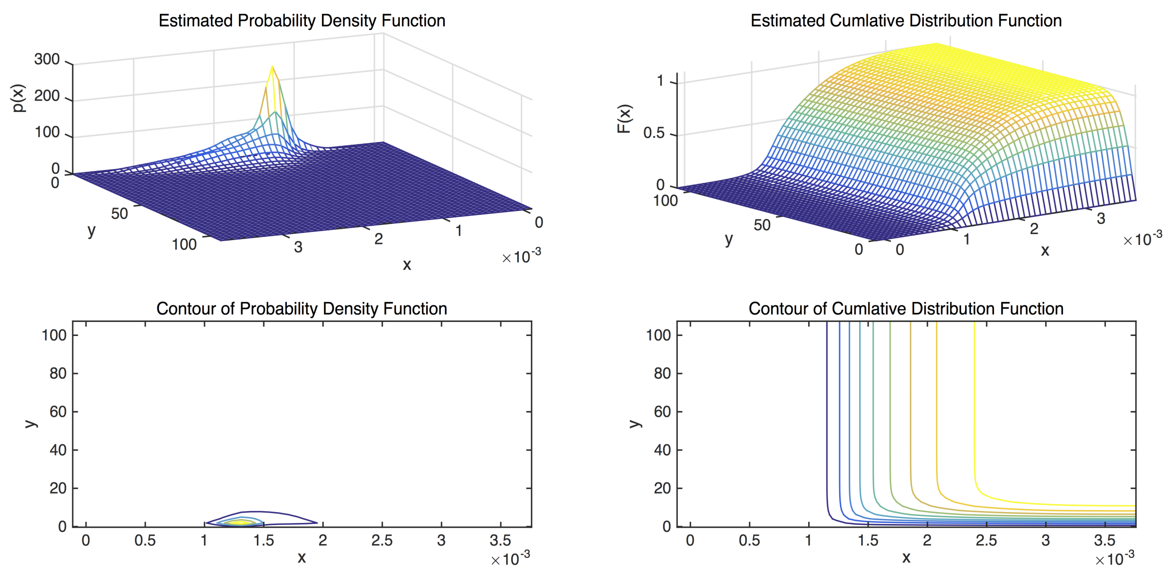
B. Extreme Index

As we can see in the four extreme transit above, the EBs and NOs is defined according to the time of travel activities, while the TIs and RIs is defined based on the amount (frequency) of travel activities. Thus, we may conclude that the extreme transit pattern is influenced by two aspects of the travel activities, time and amount. The earlier or later passengers travel, the more extreme their transit is. Also the larger the amount of trips is, the more extreme the travel pattern is. To generate a general definition of extreme transit behavior, we define the extreme level of each passenger's travel pattern as **Extreme Index** (EI) combining the travel time and travel amount. The larger the EI is, the more extreme the card holder's travel pattern is.

1) Travel Time Probability: In order to be better aware of whether the travel time of a card holder is extreme or not, we acquire the probability of each minute in the week from weekly travel time distribution shown in Figure 1(a). Taking the data of 2010 for an example, on entry of smart card record has k trips in the week, whose travel time sequence,T, is represented as $T = \{t_i | i = 1, ..., k; 0 < k < 10080\}$.



(a) PDF and CDF of 2010. The value of probability in the PDF plot is amplified to be seen clearly



(b) PDF and CDF of 2014. The value of probability in the PDF plot is amplified to be seen clearly

Fig. 4. 2-dimensional PDF and CDF of TTP and ATT of the data in 2010 and 2014

The total quantity of trips in the week is N. The number of trips at time t_i is n_{t_i}. Thus, the empirical probability of occurrence of a trip at time t_i is represented as $p_{t_i} = n_{t_i}/N$. For a smart card record, its travel time probability can be described as $TTP = \sum_{i=1}^{k} p_{t_i}/N$. Here, we get the distribution of TTP of all the SCD of 2010, shown in Figure 3. Some common-used distribution function was tried to fitting the TTP distribution, but the performance is not good. Thus, we will deal with the fitting problem in the following sections. We also explore the accuracy of the definitions of the four extreme transit behavior (EB, NO, TI and RI). The TTP's distribution of extrame travelers is compared with the TTP's distribution of whole data. Both the 2010 and 2014 cases show that the TTP's distribution of extreme travelers is smaller than that of the whole card holders, shown in Figure 5. Although, this phenomenon partially verifies the reasonability of the definitions, the extreme model still needs to be explored further.

2) Extreme Index: After the computation in the section above, each smart card has an attribute of TTP which is related with level of extreme transit behavior. The smaller TTP is, the lower probability this kind of travel time will appear with. But two card holders' travel patterns with the

TABLE II
CRITERIA OF THE DATA OF 2010 TO DETERMINE THE COMPONENT SIZE n

	2,000 samples			20,000 samples			200,000 samples			Total data (nearly 2,000,000)		
n	AIC 1.0e+4	BIC 1.0e+4	NLL 1.0e+3	AIC 1.0e+5	BIC 1.0e+5	NLL 1.0e+4	AIC 1.0e+6	BIC 1.0e+6	NLL 1.0e+5	AIC 1.0e+7	BIC 1.0e+7	NLL 1.0e+6
1	-1.2905	-1.2877	-6.4576	-1.2988	-1.2984	-6.4944	-1.2955	-1.2955	-6.4776	-1.1216	-1.1216	-5.6082
2	-1.3791	-1.3730	-6.9066	-1.3840	-1.3831	-6.9212	-1.3798	-1.3797	-6.8992	-1.2094	-1.2094	-6.0470
3	-1.4298	-1.4202	-7.1658	-1.4357	-1.4344	-7.1802	-1.4247	-1.4245	-7.1235	-1.2199	-1.2198	-6.0990
4	-1.4595	-1.4466	-7.3205	-1.4521	-1.4503	-7.2629	-1.4398	-1.4396	-7.1995	-1.2475	-1.2475	-6.2378
5	-1.4623	-1.4460	-7.3403	-1.4581	-1.4558	-7.2932	-1.4576	-1.4573	-7.2882	-1.2570	-1.2570	-6.2851
6	-1.4688	-1.4492	-7.3792	-1.4681	-1.4654	-7.3442	-1.4639	-1.4635	-7.3197	-1.2645	-1.2645	-6.3225
7	-1.4718	-1.4488	-7.4001	-1.4697	-1.4664	-7.3525	-1.4655	-1.4651	-7.3278	-1.2674	-1.2673	-6.3770

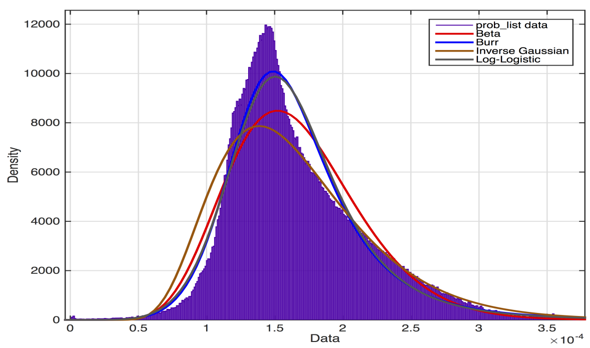
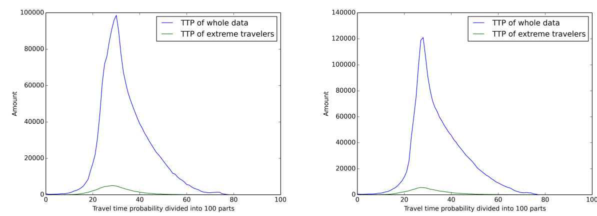


Fig. 3. Distribution of the probability of SCD's travel time probability of 2010

same amount of TTP can also be different. For example, the travel pattern with a TTP acquired by 10 travel activities is definitely different from that acquired by 2 travel activities. Thus, we estimate the traveller's EI by considering two aspects, TTP and the amount of travel times in the whole week (ATT), and acquire 2-dimensional distribution. The distribution of 2010 is shown in Figure 4(a) and the distribution of 2014 is shown in Figure 4(b). Although the two distributions of 2010 and 2014 are very similar form the view probability density function (PDF), we can see the contours of the two distribution is obviously different. The distribution of 2010 is more scattered, generating two peaks in the central area, and the distribution of 2014 is denser, concentrating in one peak. The difference between the two distribution may be lead by the higher usage of metro system, since the metro system increased 8 lines during the 4 years in Beijing.Thus, more investigations on the difference, we observed from the PDF and contour figures, deserve to be carried out.

We use Bivariate Gaussian Kernel Density Estimation to fit the TTP-ATT distribution. We also get the contour of the PDF and cumulative distribution function (CDF) which can help us have a better view of the distribution. As the 2d distribution can be represented as a mixture gaussian distribution, we fits the model by maximum likelihood using the Expectation-Maximization (EM) algorithm to get parameters. The mixture gaussian distribution can be represented by the distribution function F:



(a) 2010 (b) 2014

Fig. 5. TTP's distributions of extreme travelers and the whole card holders

$$F(x) = \sum_{i=0}^{n} \omega_i P_i(x) \tag{1}$$

As the vector $x = \{x_1, x_2\}$ obeys normal distribution:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \tag{2}$$

We use EM algorithm to estimate the n tuples of parameter (ω, μ, Σ). But the first thing we need to do is determine the value of n. There are several criteria that can evaluate the quality of the model and the goodness of fit of the model, like Akaike information criterion (AIC), Bayesian information criterion (BIC) and negative log likelihood (NLL). After we test the n ranging from 3 to 7, we get the the criteria listed in Table II. We use the three criteria mentioned above to measure the number of component of the distribution. As we can see in the table, no matter what the sample size is and no matter which criteria is, when n is larger than 4, the value of the criteria decrease very little. Especially for the case in the BIC with 2000 sample tested, when n is set to 4 (in the gray gird), the value is larger than the value when n equals 5. Thus, considering the simplicity of a model and the goodness of the fit, we choose n equals 4 in this model. Figure **??** shows the contour of our mixture gaussian distribution model. For the 2010 case, the parameters with 4 tuples (ω, μ, Σ) of the model is described as:$\{((\omega_1 = 0.327), (\omega_2 = 0.198), (\omega_3 = 0.161), (\omega_4 = 0.314)); (\mu_1 = (0.0019, 10.24), (\mu_2 = (0.0013, 1.46), (\mu_3 = (0.0020, 2.57), (\mu_4 = (0.0014, 5.47)); ((\Sigma_1 = \left(\begin{smallmatrix} 0.0000 & -0.0009 \\ -0.0009 & 24.8318 \end{smallmatrix}\right), (\Sigma_2 = \left(\begin{smallmatrix} 0.0000 & 0.0000 \\ 0.0000 & 0.4341 \end{smallmatrix}\right), (\Sigma_3 = \left(\begin{smallmatrix} 0.0000 & -0.0002 \\ -0.0002 & 1.6797 \end{smallmatrix}\right), (\Sigma_4 = \left(\begin{smallmatrix} 0.0000 & 0.0001 \\ 0.0001 & 6.0962 \end{smallmatrix}\right)))\}$. The parameters of 2014 is described

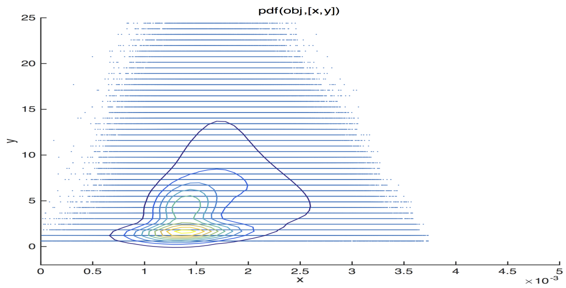


Fig. 6. Estimated contour of the mixture gaussian distribution of 2010 data. X-axis represents the TTP and y-axis represents the ATT

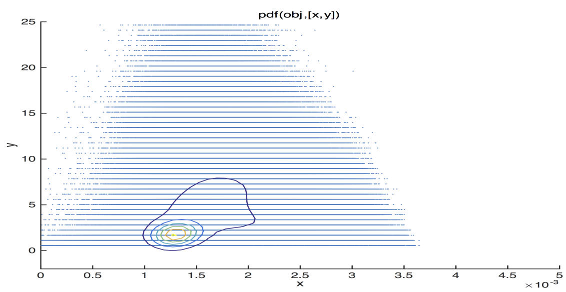


Fig. 7. Estimated contour of the mixture gaussian distribution of 2014 data. X-axis represents the TTP and y-axis represents the ATT

as:$\{((\omega_1 = 0.190),(\omega_2 = 0.358),(\omega_3 = 0.138),(\omega_4 = 0.314));(\mu_1 = (0.0017,10.48),(\mu_2 = (0.0021,5.52),(\mu_3 = (0.0013,1.33),(\mu_4 = (0.0014,3.12));((\Sigma_1 = \left(\begin{smallmatrix}0.0000 & 0.0001\\ 0.0001 & 24.5545\end{smallmatrix}\right),(\Sigma_2 = \left(\begin{smallmatrix}0.0000 & -0.0006\\ -0.0006 & 8.9095\end{smallmatrix}\right),(\Sigma_3 = \left(\begin{smallmatrix}0.0000 & 0.0000\\ 0.0000 & 0.3349\end{smallmatrix}\right),(\Sigma_4 = \left(\begin{smallmatrix}0.0000 & 0.0001\\ 0.0001 & 2.3395\end{smallmatrix}\right)))\}$.

Thus, up to now, we have the model of EI and let EI equals the distribution function shown the equations in the Eq. 1 and 2. Given a SCD, we can quickly compute the TTP and ATT as the input variables to get the value of EI. The smaller the EI, the more extreme the transit pattern is.

The EI models of 2010 and 2014 seems similar according to their parameter tuples. But as we can see in the Figure 6 and 7, which show the estimated contour of the PDF in 2010 and 2014, the distribution are actually different. Especially for the cohesion, the EI model of 2014 is more simple and regular-shaped. We also find that the average EI of the four extreme pattern is lower that the average of the whole dataset mentioned above, which can in return prove the legitimacy of the EI model.

VI. Conclusion

In this paper, we focus on the evolution of extreme public transit behaviors of travelers by using the SCD of Beijing in the year of 2010 and 2014. We carry out several descriptive statistics of the SCD in order to have a better understanding the dynamic process and evolution of the extreme transit behavior. These distributions based on the SCD's temporal information reveals several interesting phenomena and corroborate each other with the datasets' background. By combining the SCD's temporal information and the amount of travel behavior, we proposed a concept of Extreme Index (EI) based on the mixture Gaussian model to depict the extreme level of the passengers' travel pattern.

According to our analysis, the normal transit behavior of the two years have nearly the same temporal distributions. By analyze the distribution of TTP, we found the previous definition of four extreme transit behavior is reasonable. The EI models of two years, which are the main contribution of this paper, have similar distribution. But the EI model of 2010 with two peaks is more scattered than that of 2014, which has only one peak. The difference between the two EI model may be lead by the higher usage of metro system, since the metro system increased 8 lines during the 4 years in Beijing. With the EI model, every smart card will have an attribute of EI, which can be introduced to further cross comparisons analysis. We will also improve the EI model with respect to the variation of trips' temporal information and entropy of the occurrence of travel activities. The EI model can definitely be applied in further discussion for urban transit and passengers' behavior and status analysis.

VII. Acknowledgments

The authors would like to appreciate the partial funding support from National Natural Science Foundation of China (No. 51329801 and 51408039).

References

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

[2] Z. Cui and Y. Long. Perspectives on stability and mobility of passenger's travel behavior through smart card data. In *ACM SIGKDD Workshop on Urban Computing*, page 9p, 2015.

[3] M. Dou, T. He, H. Yin, X. Zhou, Z. Chen, and B. Luo. Predicting passengers in public transportation using smart card data. In *Databases Theory and Applications*, pages 28–40. Springer, 2015.

[4] 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.

[5] 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.

[6] E. F. Legara and C. Monterola. Inferring Passenger Type from Commuter Eigentravel Matrices. *arXiv:1509.01199 [physics, stat]*, Aug. 2015. arXiv: 1509.01199.

[7] 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.

[8] Y. Long and J.-C. Thill. Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing (in press). *Computers, Environment and Urban Systems*, 2015.

[9] 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.

[10] 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.

[11] 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.