Urban Population Migration Pattern Mining Based on Taxi Trajectories

Bing Zhu
Tsinghua University
Beijing, China
zhub.daisy@gmail.com

Qixing Huang
Stanford University
Stanford, CA, U.S.A.
huangqx@stanford.edu

Leonidas Guibas
Stanford University
Stanford, CA, U.S.A.
guibas@cs.stanford.edu

Lin Zhang
Tsinghua University
Beijing, China
linzhang@tsinghua.edu.cn

ABSTRACT
Understanding urban population migration patterns is very helpful for urban operation and management, including the traffic forecasting, epidemic prevention, commercial resource allocation, emergency response and future urban planning. The large taxi fleet equipped with GPS comprises ubiquitous mobile probes in urban areas, and their trajectories reveal interesting phenomena in the city. We investigate the urban population migration pattern mining based on taxi trajectories, including 1) the hotzone identification and 2) the principal Origin-Destination traffic flow (OD pair) extraction. We show how to apply state-of-art point clustering algorithms to address these two tasks. The performance of the algorithms are evaluated on a Beijing taxi trajectory dataset generated by 8,000 taxicabs in May 2009. The results show interesting insights of the time-resolved results, which is consistent with semantic interpretations.

Keywords
ubiquitous computing, GPS trajectory, population migration pattern, density-based clustering

1. INTRODUCTION
Taxis, as one of the most important and widely used public transportation methods, can tell lots of stories about the city. In modern cities, large taxi fleets are always equipped with GPS, turning them into ubiquitous mobile probes of the urban areas. For instance, there are 66,600 licensed taxicabs in Beijing in 2012, which generate over 1.2 million ridden trips per day. The taxi fleet also provides a very good coverage of the city in both space and time. Driven by drivers’ profit-seeking objective and passengers’ demands, impacted by the traffic pattern on road network, the taxi trajectories can provide great insights into the urban hotspot distribution, population migration pattern, traffic pattern and road network, which will serve the Location Based Service (LBS) applications, transportation system management, and urban planning.

We investigate the pattern mining of urban dynamics from taxi trajectories from two aspects. One is to identify the hotzones that indicate the high level of human activities, indicated by large number of taxi pickups and drop-offs. The other is to explore the principal OD flows that reveal the resident migration pattern. Our work is motivated from the foundation work of [4] by Gonzalez, et al. They found that human trajectories showed a high degree of temporal and spatial regularity.

We formulate both the hotzone identification and the principal OD pair extraction as the spatial-temporal point clustering problem and apply density-based clustering techniques to solve them. The contribution of this paper lies in two aspects:

- **Spatial-temporal hotzone identification**: We modify the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to identify the time-resolved hotzones, and apply it to the picking-up and dropping-off location point clouds to identify hotzones in Beijing.

- **OD pair clustering**: We define a parametric distance metrics to measure the difference between two OD pairs. We apply the DBSCAN method to reveal the characteristic travel patterns in the city.
2. MINING URBAN POPULATION MIGRATION PATTERN

In this section, we describe the technical aspect of our methods for mining population migration patterns using taxi trips. Suppose we are given $N$ trips $\mathcal{T} = \{t_1, \ldots, t_N\}$. Each trip $t_i$ is represented as an original-destination pair (OD pair) $O_i, D_i$, where $O_i$ and $D_i$ denote the original (pick-up) and the destination (drop-off) locations, respectively. Each location $O_i$ (or $D_i$) is given by three coordinates, i.e., normalized longitude $O_i^x$ (or $D_i^x$), latitude $O_i^y$ (or $D_i^y$) and time sample $O_i^t$ (or $D_i^t$). To make the notations uncluttered, we will always use $\|\cdot\|$ to denote the 2D Euclidean distance between two locations, e.g., $\|O_i - D_i\| = ((O_i^x - D_i^x)^2 + (O_i^y - D_i^y)^2)^{1/2}$.

As mentioned above, we consider two clustering problems for mining migration patterns. In the first problem, we identify time-dependent hotzones by clustering the collection of pick-up and drop-off locations of all input trips. In the second problem, we identify the flow pattern by clustering the OD pairs. For both clustering problems, we employ the DBSCAN algorithm described by Ester et al. [2], which has proven to be successful on large datasets [11]. In the reminder of this section, we first present an overview of the DBSCAN algorithm. Then we show how to apply it to solve the two proposed clustering problems.

2.1 DBSCAN Clustering

The DBSCAN clustering algorithm proposed by Ester et al. [2] takes as input a collection of points $\mathcal{D} = \{p\}$ and a pre-defined distance measure $d(\cdot, \cdot): \mathcal{D} \times \mathcal{D} \to R$ between pairs of points, and outputs a decomposition of $\mathcal{D} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_m \cup \mathcal{N}$, such that points in each cluster $\mathcal{C}_i$ are close to each other in terms of the pre-defined distance measure, and isolated points are categorized into noise set $\mathcal{N}$.

The DBSCAN algorithm is controlled by two parameters $\epsilon_d$ and $n_{\text{min}}$. $\epsilon_d$ defines the neighbor of each $p \in \mathcal{D}$: $\mathcal{N}(p) = \{p' | d(p, p') < \epsilon_d\}$, and $n_{\text{min}}$ is used to classify a point $p$ into a border point or a core point based on the size of its neighbor set, i.e., $p$ is a core point if and only if $|\mathcal{N}(p)| > n_{\text{min}}$.

Given $\epsilon_d$ and $n_{\text{min}}$, the algorithm builds an oriented adjacency graph $\mathcal{G}$ over $\mathcal{D}$. An edge $(p, p') \in \mathcal{G}$ if and only if $p$ is a core-point and $p' \in \mathcal{N}(p)$. The clusters $\mathcal{C}_i$ are simply given by the the connected components of $\mathcal{G}$. Note that clusters, which contain a single point, are removed from the output, i.e., they are considered noise.

We choose the DBSCAN algorithm for our clustering purposes due to its simplicity. In particular, it does not require the number of clusters as input. Moreover, it can handle clusters with arbitrary shape.
2.2 Hotzone Extraction

To extract hotzones, we feed the pick-up and drop-off locations of all input trips into the DBSCAN algorithm. In this case, the input point set \( D = \{ O_i, 1 \leq i \leq n \} \cup \{ D_i, 1 \leq i \leq n \} \). As there is no obvious scaling factor between the time sample and the longitude (latitude), we add another threshold \( \epsilon_t \) to determine the neighbors of each point \( p \), i.e., \( N(p) = \{ p' \| p - p' \| < \epsilon_t, |p' - p'| < \epsilon_t \} \). We will analyze the clustering results in Section 3.

2.3 Origin-Destination traffic flow clustering

To obtain clusters of original-destination traffic flows, we consider each OD pair \( (O_i, D_i) \) a point \( (O_i^x, D_i^y, O_i^y, D_i^y) \) in \( \mathbb{R}^4 \). When applying the DBSCAN algorithm, we found that a meaningful distance measure between OD pairs is central to the meaningfulness of the resulting clusters. Motivated from the needs to discover both long and short characteristic migration patterns and by the fact that for clusters of OD pairs with similar purposes, clusters consist of longer OD pairs tend to be more diverse than those of shorter ones, we define the distance measure between two OD pairs \( O_i D_i \) and \( O_j D_j \) as:

\[
d(\alpha)(O_i D_i, O_j D_j) = \frac{\left\| \|O_i - O_j\|^2 + \|D_i - D_j\|^2 \right\|^{\frac{\alpha}{2}}}{\left(\|O_i - D_i\| + \|O_j - D_j\|\right)^\alpha}
\]

(1)

where \( \alpha \geq 0 \) is a normalization parameter. In other words, \( d(\cdot, \cdot) \) is a relative distance, which is normalized by the length of the OD pairs.

The normalization parameter \( \alpha \) adjusts the length of the OD pairs in the resulting clusters. When \( \alpha = 0 \), the distance metric degenerates to the absolute Euclidean distances between OD pairs. In this case, long OD pairs are unlikely to be clusters. In contrast, when \( \alpha \) becomes larger, we allow bigger absolute distances between two long OD pairs, and thus they are more preferred to be clustered.

To prevent the chain phenomenon in clustering, we check the bounding boxes of the origins and destinations of each OD pair cluster \( C_i \). If the area of one of the bounding boxes is bigger than \( a_{max} \), we then apply the DBSCAN algorithm to further subdivide \( C_i \). To ensure that \( C_i \) is subdivided into multiple clusters, we reduce the distance threshold by half, i.e., \( \epsilon_t = \epsilon_t/2 \). This process is iterated until all clusters satisfy the area constraints.

3. EVALUATION

In this section, we carry out our methods on a large-scale GPS trajectory dataset and evaluate their performances.

3.1 Dataset

The trajectory dataset was generated by approximately 8,000 taxicabs in Beijing in May, 2009. We process the log file of each taxicab to extract individual trips. There are about 120k trips generated in one day.

3.2 Urban hotzone identification

We applied out hotzone extraction method on the trips whose staring time is between 8:00 a.m. and 9:00 a.m. on May 1st 2009. As shown in Fig. 3, our algorithm identifies 37 clusters. Projecting the clusters onto the map, we can see that the locations of these clusters and their shapes matched perfectly to certain urban functional areas in Beijing including the Capital International Airport terminals, the railway stations, the public transportation center (Xizhimen), the Central Business District (CBD) and the university town. By comparing this result with those achieved from the trip data between 11:00 and 12:00 pm, as shown in Fig. 3, in which only 19 clusters in the residential areas and transportation centers are grouped, one can show that the algorithm reveals an obviously less active status in the late night.

3.3 OD pair clustering

Figure 2 shows the results of the OD pair clustering algorithm on the trips from 8:00 to 9:00 a.m., May 1st, 2009. Here we tested three different values of \( \alpha = 0, 0.5 \) and 1. As illustrated Figure 2, the OD pair clustering algorithm is able to identify popular flows such as between the airport and the downtown area and between large residential area and the business districts.

\( \dagger \)The distance threshold is further reduced by half if this procedure is applied on the sub-clusters of \( C_i \).
Figure 3: (a) Hotzones (11:00-12:00pm, May 1st, 2009), (b) OD pair clusters of the same period.

Adjusting $\alpha$ yields different types of clusters. When $\alpha = 0$, the algorithm only considers absolute distance and finds clusters that are more concentrated, e.g., from the airport to the closest subway station. In contrast, when $\alpha = 1$, the algorithm tolerates more errors for long trip and outputs clusters that composed of more scattered long trips, e.g., clusters of trips from the airport to the union of large tourist zones.

We then applied the OD pair clustering algorithm on trips of different time period. The following are some interesting results and their semantic interpretations, which justify the effectiveness of the OD pair clustering algorithm. As illustrated in Figure 3 and Table 1, we observe that more OD pair clusters are found during the daytime than early morning and late night, because more pick-up events occur and more OD pair clusters are grouped. Second, higher ratio of OD pair are included into the clusters during the active hours, so we can infer that people’s moving behaviors are more similar and converge to the principal OD pairs, especially the ones involving business districts and transportation centers, which brings great opportunities and challenges for the future city planning. Third, apart from the level of activity corresponding to the traffic volume, in different time periods through a day, specific traffic patterns vary. For example, in the early morning (Fig. 3(a)), the main flows are those connecting the residential areas and transportation centers, while at night, a large cluster from CBD to the suburban residential area (the red cluster in the right side of Fig. 3(b)) emerges. This cluster doesn’t show up before 11:00 p.m., when the subway Line 1 stops service, and this is consistent to the fact that subway is a preferable transportation method to taxi in Beijing and also implies the subway service should probably be extended in time.

Different patterns also can be found between weekdays and holidays. In contrast to the workday afternoon traffic pattern on May 25th, 2009 (Fig. 3(c)), people behave much more actively in the holiday afternoons such as Fig. 3(b) on May 1st, a national holiday.

Comparing to the hotzone clustering, the high-order OD flow clustering algorithm reveals more details. Based on the data from 11:00 to 12:00pm, in Fig. 3(b) the destinations of the blue flow cluster from the airport and the red one from the Wangfujing Shopping Street are distinguished, while they both fall into the green hotzone between Dongdan and Jianguomen in Fig. 3(a).

The blue destination area turn out to be some hotels included into the clusters during the active hours, so we can infer that people’s moving behaviors are more similar and converge to the principal OD pairs, especially the ones involving business districts and transportation centers, which brings great opportunities and challenges for the future city planning. Third, apart from the level of activity corresponding to the traffic volume, in different time periods through a day, specific traffic patterns vary. For example, in the early morning (Fig. 3(a)), the main flows are those connecting the residential areas and transportation centers, while at night, a large cluster from CBD to the suburban residential area (the red cluster in the right side of Fig. 3(b)) emerges. This cluster doesn’t show up before 11:00 p.m., when the subway Line 1 stops service, and this is consistent to the fact that subway is a preferable transportation method to taxi in Beijing and also implies the subway service should probably be extended in time.

### 4. CONCLUSION

In this paper, we investigate the urban population migration pattern based on the taxi trips. Focusing on the pick-up and drop-off events, the urban hotzone identification and principal traffic flow extraction problems are solved by the spatial-temporal density-based clustering method we proposed. Specifically, the principal traffic flow analysis is formulated as a 4-D point clustering problem and the relative distance between two OD pairs is defined, including a preference factor which can be used to fine-tune the preference to cluster length. Experiments on the Beijing taxi trajectory dataset have been conducted to evaluate the performances of the methods we proposed. The trajectory dataset was generated by 8,000 Beijing taxicabs in May 2009. The results show that, first the urban hotzones can be identified, whose locations and shapes matched satisfactorily to the functional areas in Beijing. Second,
the main traffic flow clusters are successfully grouped by the 4-D point clustering algorithm and the clustering preference factor does have a great control over the clustering results. Third, combining the time-dependent clustering results with Beijing’s geographic-social background information, some insights underlying the results are interpreted, which can justify the effectiveness of the method.

5. REFERENCES


