2016 17th IEEE International Conference on Mobile Data Management

Understanding Urban Mobility via Taxi Trip Clustering

Dheeraj Kumar[1], Huayu Wu[2], Yu Lu[2], Shonali Krishnaswamy[2], Marimuthu Palaniswami[1]

[1]Electrical and Electronic Engineering, The University of Melbourne, Australia
[2]Data Analytics Department, Institute for Infocomm Research (I2R),
Agency for Science, Technology and Research (A*STAR), Singapore

{dheerajk@student., palani@} unimelb.edu.au, {huwu, luyu, spkrishna}@i2r.a-star.edu.sg

***Abstract*—Clustering of a large amount of taxi GPS mobility data helps to understand the spatio-temporal dynamics for the applications of urban planning and transportation. In this paper we cluster the origin-destination pairs of the passenger taxi rides to provide useful insight into the city mobility patterns, urban hot-spots, road network usage and general patterns of the crowd movement within the city of Singapore. We perform experiments on a large scale Singapore taxi dataset consisting of more than 10 million passenger origin-destination GPS points. We use the clusiVAT sampling scheme to obtain the sample trips which return coarse clusters describing the major crowd movement and reduce the data points that are not captured by the coarse clusters and may bring in noises during fine-grained clustering. After the sampling step we use the well known density based clustering algorithm DBSCAN to find cluster structure in the sampled datapoints and later extend it to the rest of the dataset using nearest prototype rule. We report 24 trip clusters from the dataset which are compact enough to draw meaningful conclusions about the city mobility patterns and the number of trips in each cluster is large enough to be representative of the general traffic movement.**

I. Introduction

Advent of GPS enabled devices and mobile phones have made mobility data readily available. Analysis of such a large amount of mobility data can help understand the space time dynamics for the applications of urban planning and transportation. Many cities around the world have introduced the concept of GPS sensor equipped taxis to enable taxi on call and taxi tracking services. Apart from catering to these services, these devices generate a rich GPS trace mobility data which could provide useful insight into the city mobility patterns, urban hot-spots, road network usage, etc. Although GPS sensor equipped public transport such as buses and trams is quite a norm, they have limited coverage spanning fixed routes only. Taxis on the other hand have a wide coverage and in a city like Singapore are one of the major mode of transportation for the public, hence providing a fair estimate of general mobility trends of people and of city hotspots. Depending on the data characteristics (e.g. spatial and temporal resolution) and the research question to be addressed, the data analytics could be performed on the taxi trajectories (a sequence of GPS points sampled at a regular interval) or on the origin-destination pair for each passenger ride. In this paper we focus on the latter class, the analysis of the origin-destination GPS pair for taxi rides as the entire trajectories cannot reflect the boarding and alighting pattern, and thus cannot reflect the crowd movement. Specifically, we use data clustering to obtain general patterns of the crowd movement within the city of Singapore.

Clustering of origin-destination locations of passenger taxi trips could provide useful insight into the passenger movement pattern across the city. It also helps in identifying the hot-spot locations in the city where the taxi drivers are most likely to find their next customer. In this paper we perform clustering analysis on the origin-destination GPS location of the taxi trips in Singapore. A review of major data clustering approaches and their applications can be found in [1]. Based on the characteristics of the dataset to be experimented on and the desired outcome, we choose the best possible clustering technique for our purpose as described below. The presence of a large number of noise points, no guarantee regarding the convex shape of the clusters and no priori information about the number of clusters, k, to seek make partitional clustering algorithm such as k-means unsuitable for clustering this type of dataset. The hierarchical and density based schemes, although does not require k as input and can find clusters of any random shape, they becomes computationally prohibitive for such a huge dataset. In such cases, usually random samples are picked from the large dataset before applying density or hierarchical clustering schemes, but the random samples may not be representative of the entire dataset and we may need to perform the experiment multiple times to obtain all the clusters. In this paper we use the sampling scheme proposed in [2]–[4] to obtain samples which return coarse clusters describing the major crowd movement trends and minimize the datapoints which are not in the coarse clusters as some clusters may cover a very broad region due to the dense and wide distribution of data points in such an area. After the sampling step we use a density based clustering scheme DBSCAN [5] to find cluster structure in the sampled datapoints and later extend it to the rest of the dataset using nearest prototype rule.

The rest of the paper is organized as follows. Section II provides a brief literature survey and describes the related work being done in this field. Section III provides the detail description of the Singapore taxi GPS log dataset and the procedure to extract taxi trips from the log dataset. In section IV, we describes the proposed clustering scheme including the sampling algorithm. The results of the experiments performed

2375-0324/16 $31.00 © 2016 IEEE
DOI 10.1109/MDM.2016.54

TABLE I. TAXI STATUS DESCRIPTION

BUSY	Taxi driver temporarily unavailable due to a personal reason
STC	Taxi soon to clear the current job and ready for new bookings
FREE	Taxi unoccupied and ready for taking new passengers or bookings
BREAK	Taxi on a break and driver logged on mobile data terminal (MDT)
POWEROFF	MDT shut down and not working
ARRIVED	Taxi arrived at the booking pickup location and waiting for the passenger
ONCALL	Taxi unoccupied, but accepted a new booking job
OFFLINE	Taxi on a break and driver logged off from MDT
POB	Passenger on board and taximeter running
PAYMENT	Passenger making payment and taximeter paused
NOSHOW	No passenger showing up and the booking canceled soon

on the Singapore taxi datatset are given in section V before concluding in section VI.

II. RELATED WORK

Taxi origin-destination pair location clustering to discover and understand spatio-temporal patterns in movement is a relatively new and exciting field of research. Yamamoto et al. [6] proposed an adaptive routing method for the cruising taxis by suggesting vacant taxis to the pathways having many potential passengers using a clustering approach. Authors in [7], [8] used data mining techniques such as clustering and naive Bayesian classifier on historical data to build models and predict taxi demand in contexts of time, weather, and location. However the works described above focus more on analyzing the taxi trajectory as a whole rather than identifying the origin destination clusters for taxi trips.

Wan et al. [9] used a density based hierarchical clustering method they called "DBH-CLUS" to identify pick-up/drop-off hotspots to propose better locations for setting up taxi stands. Guo et al. [10] presented a new approach to the discovery and understanding of spatio-temporal patterns in the passenger movement using spatial clustering of the origin destination GPS points to recognize potentially meaningful places and map the clusters to understand the spatial distribution and temporal trends of movements. Other notable work which perform origin-destination GPS location clustering on taxi trajectory data to interpret urban hotspots and crowd movement patters include [11]–[14].

Data clustering is the problem of partitioning a set of unlabeled objects $O = \{o_1, o_2, ..., o_n\}$ into k groups of similar objects, where $1 < k < n$. Most clustering algorithms can be divided into following four categories:

- Connectivity based clustering (hierarchical clustering)
- Centroid-based clustering
- Distribution-based clustering
- Density-based clustering

Hierarchical clustering relies on the fact that nearby objects have a higher probability of belonging to the same cluster than to a cluster containing objects that are farther away. This category includes single linkage (SL), which is based on cutting large edges in a minimum spanning tree (MST) [15]. The hierarchical clustering approaches are most general and can be applied to any dataset (vector or relational), however have high computational complexity and suffer from adverse effect of *Chaining* [16] in presence of noise. Centroid-based algorithms such as k-means [17], [18] depends on optimizing an objective function, which typically measures a property such as inter-cluster separation, within-cluster variance or both. Although easy to implement and being computationally efficient, the k-means algorithm require k as an input which is usually not known. Another limitation of k-means is that it tries to impose elliptical shape on all k clusters and hence is not suitable for discovering oddly shaped clusters. Distribution based clustering approaches such as Gaussian Mixture Model (GMM) consider the data to be generated from a Gaussian mixture model and try to find model parameters which are most likely to produce this dataset. Although theoretically sound, these methods suffer from a common problem called *overfitting*. Since a more complex model will usually be able to explain the data better, choosing the appropriate model complexity becomes difficult. Density based clustering schemes such as DBSCAN [5] and OPTICS [19] define clusters as areas of higher density than the remainder of the data set. These schemes does not require the number of clusters, k to seek as an input and can find clusters of any arbitrary shape but become computationally prohibitive as the dataset size increases due to computational complexity of $O(n^2)$ for calculating the distance matrix.

III. TAXI DATA AND TRIP EXTRACTION

A. Data

The dataset consists of the trajectories of more than $15,000$ taxis collected over a duration of 1 month. The dataset is very dense as it consists of more than 370 million datapoints. The general format of each datapoint is as follows: {Time Stamp, Taxi Registration, Latitude, Longitude, Speed, Status}. The status field of each datapoint consists of one of the 11 values as described in Table I.

This GPS log dataset is processed to obtain taxi trip information before applying the clustering framework on the origin-destination GPS location of the taxi trips. The trip extraction procedure is described in the following subsection.

B. Trip Extraction

We firstly define several important terms used in the trip extraction.

Definition 1. Individual taxi raw trajectory $\Re$: A temporally ordered sequence of the taxi raw data records from one taxi, i.e., $p_1 \rightarrow \cdots \rightarrow p_i \rightarrow \cdots \rightarrow p_n$, where p_i $(1 \leq i \leq n)$ is the tuple containing the taxi state $p_{i.state}$, instantaneous speed $p_{i.speed}$, latitude coordinate $p_{i.lat}$, longitude coordinate $p_{i.lon}$ and timestamp $p_{i.ts}$.

Definition 2. Taxi single trip $R(s, e)$: A temporally ordered sequence of taxi's raw data records for a single trip, i.e., $p_s \rightarrow p_{s+1} \rightarrow \cdots \rightarrow p_e$, where $1 \leq s < e \leq n$.

Definition 3. Taxi trip set ω: A collection of taxi's trip trajectory, i.e., $\{R^k | k = 1, 2, \cdots\}$, where $R^k = R(s_k, e_k)$.

Definition 4.1 Taxi occupied state set Θ: { POB, STC, PAYMENT }.

Definition 4.2 Taxi unoccupied state set Ψ: { FREE, ONCALL, ARRIVED, NOSHOW }.

Definition 4.3 Taxi non-operational state set Λ: { BREAK, BUSY, OFFLINE, POWEROFF }.

In order to extract each individual taxi's trips from the raw taxi data, we propose a simple and practical algorithm, called trip extraction algorithm (TEA): its input is an individual taxi's raw trajectory $\Re$ and output is the taxi trip set ω. The complete algorithm is shown in Algorithm 1. The basic idea behind the TEA algorithm is that a complete trip normally consists of a certain taxi state transitions, e.g., starting from *FREE* to *POB* and ending from *POB* to *FREE*. The algorithm uses a flag δ to mark whether a trip starts or not, adds the each new R^k into the trip set ω, given R^k satisfies the state transition constraint: R^k starts with an occupied state and end with an unoccupied state or non-operational state. In the practical implementation, the system also filters out the trips with too short duration or multiple non-operational states.

Algorithm 1: Trip Extraction Algorithm

Input : Taxi trajectory $\Re$.
Output: The extracted trip set ω

$\delta \leftarrow false$; $k \leftarrow 1$;
for $i \leftarrow 1$ **to** $length(\Re)$ **do**
  if $p_{i.state} \in \Theta$ *and* δ=*false* **then**
    $R^k.Add(p_{i-1})$;$R^k.Add(p_i)$;$\delta \leftarrow true$
  else
    if $p_{i.state} \in \Theta$ *and* δ=*true* **then**
      $R^k.Add(p_i)$
    else
      if $p_{i.state} \in \Psi \cup \Lambda$ *and* δ=*true* **then**
        $R^k.Add(p_i)$
        $\omega.Add(R^k)$
      else
        $k \leftarrow k + 1$;$\delta \leftarrow false$
      end
    end
  end
end

IV. The proposed clustering scheme

The clustering scheme we use for clustering origin destination GPS datapoint for taxi trips consists of two stages as described below:

A. The sampling step

Instead of using the random sampling, which is generally used for large datasets, we use the sampling scheme proposed in our previous work [2]–[4]. We call this as clusiVAT sampling, which extracts the coarse clusters which describe the major crowd movement and samples those datapoints which truly represent the structure of the dataset in the p dimensional space, where p is the cardinality of each datapoint.

Consider a dataset $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ consisting of N p-dimensional datapoints where N is large (of the order of 10^6). The sampling scheme take two more inputs other than the big dataset X. These are k', which is an overestimate of number of clusters in the dataset and n, the number of samples to find. The first step for sampling is the selection of k' distinguished objects which are at a maximum distance from each other. This step divides the entire dataset into k' partitions which (on average) span almost equally sized subspaces of $\mathbb{R}^p$. The next step in clusiVAT sampling is to randomly select objects from the k' partitions to get a total of n samples. The number of objects selected from each partition is proportional to the number of datapoints in that partition. These n samples, which are just a small fraction of N, retain the approximate geometry of the dataset. The pseudocode for the sampling step is given in Algorithm 2.

Algorithm 2: clusiVAT sampling [2]–[4]

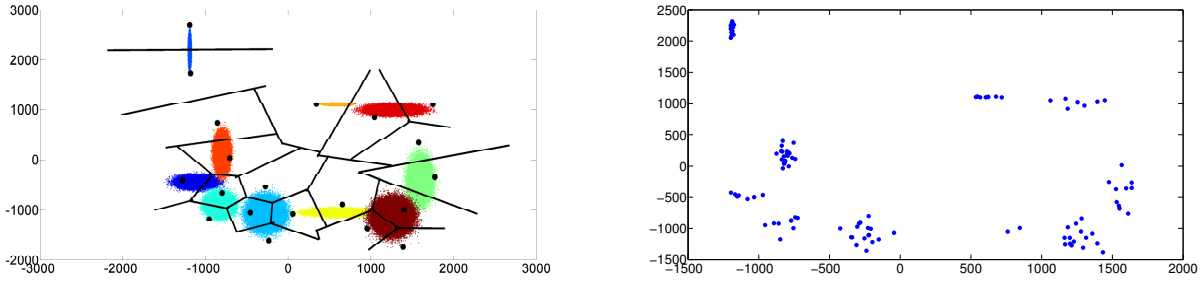
Input : $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ $-$ N p-dimensional data points
k' $-$ cluster number overestimate
n $-$ approximating sample size
Output: X_n $-$ Sampled dataset

Select the indices m **of** k' **distinguished objects**
$m_1 = 1$;$y = \{dist\{\mathbf{x_1}, \mathbf{x_1}\}, ..., dist\{\mathbf{x_1}, \mathbf{x_N}\}\}$
for $t \leftarrow 2$ **to** k' **do**
  $y = (min\{y_1, dist\{\mathbf{x_{m_{t-1}}}, \mathbf{x_1}\}\}, ...,$
  $min\{y_N, dist\{\mathbf{x_{m_{t-1}}}, \mathbf{x_N}\}\})$
  m_t=arg $\max_{1 \leq j \leq N} \{y_j\}$
end

Group objects in $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ **with their nearest distinguished objects**
$S_1 = S_2 = ... = S_{k'} = \emptyset$
for $t \leftarrow 1$ **to** N **do**
  $l = \arg \min_{1 \leq j \leq k'} \{dist\{\mathbf{x_{m_j}}, \mathbf{x_t}\}\}$;$S_l = S_l \cup \{t\}$
end

Randomly select data near each distinguished object to form X_n
for $t \leftarrow 1$ **to** k' **do**
  $n_t = \lceil \frac{n \times |S_t|}{N} \rceil$
  Draw n_t unique random indices $\widetilde{S}_t$ from S_t
end
$\widetilde{S} = \bigcup_{t=1}^{k'} \widetilde{S}_t$;$X_n = X_{\widetilde{S}}$

To illustrate the sampling procedure using an example, consider a 2-dimensional dataset shown in Fig. 1(a). It consists of $k = 10$ clusters comprising 1,000,000 points, which are intermixed with each other and hence difficult to cluster for any algorithm. In this experiment we use $k' = 20$ and $n = 100$.



(a) Ground truth scatter plot (Different colors representing different cluster)

(b) Random samples from each partition

Fig. 1. Distinguished object and random sample selection for clusiVAT sampling

Fig. 1(a) also shows the 20 distinguished objects found using the clusiVAT sampling algorithm (shown by bold black dots) and their corresponding partition of the dataset (shown by solid black lines). View (b) shows 100 randomly chosen samples.

Since the Singapore taxi dataset is quiet dense and have a lot of datapoints which are quiet close to each other and consists of a lot of noise points we refrain from using a single linkage based hierarchical clustering such as VAT [20] as it is prone to chaining effect and does not produce good results. Instead we use density based clustering DBSCAN on the sample datapoints obtained using clusiVAT sampling.

B. *Fine-grained clustering*

After the sampling step we perform fine grained clustering on the sample trips using the well known density based clustering algorithm, DBSCAN [5]. It is a locality-based clustering algorithm which assumes that points inside clusters distribute randomly but do not require k, the number of clusters. It also has a notion of noise data and is robust to outliers. DBSCAN requires two input parameters:

1) Neighbourhood considered as neighbouring region (ϵ)
2) Minimum number of points required to form a dense region (minPts)

It starts with an arbitrary starting point, and if its ϵ-neighborhood contains atleast minPts points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ϵ-environment of a different point and hence be made part of a cluster. If a point is declared to belong to a cluster, its ϵ-neighborhood is also part of that cluster. This process continues until the density-connected cluster is completely found.

V. Experiments

We perform experiments on large scale Singapore taxi data. Using the trip extraction algorithm as described in Section III-B, we extract more than 10 million passenger origin-destination pairs on which clustering is performed. Each taxi trip is represented by a pair of GPS coordinates (latitude and longitude). So each taxi trip x_i is represented by a four dimensional vector $x_i = \{Lat_{o_i}, Lon_{o_i}, Lat_{d_i}, Lon_{d_i}\}$. The distance measure between a pair of origin-destination passenger trips is defined as the sum of distance between origins and destinations respectively of the trips (in kilometers). Consider two taxi trips $x_i = \{Lat_{o_i}, Lon_{o_i}, Lat_{d_i}, Lon_{d_i}\}$ and $x_j = \{Lat_{o_j}, Lon_{o_j}, Lat_{d_j}, Lon_{d_j}\}$, the distance between them, D_{ij} is defined as

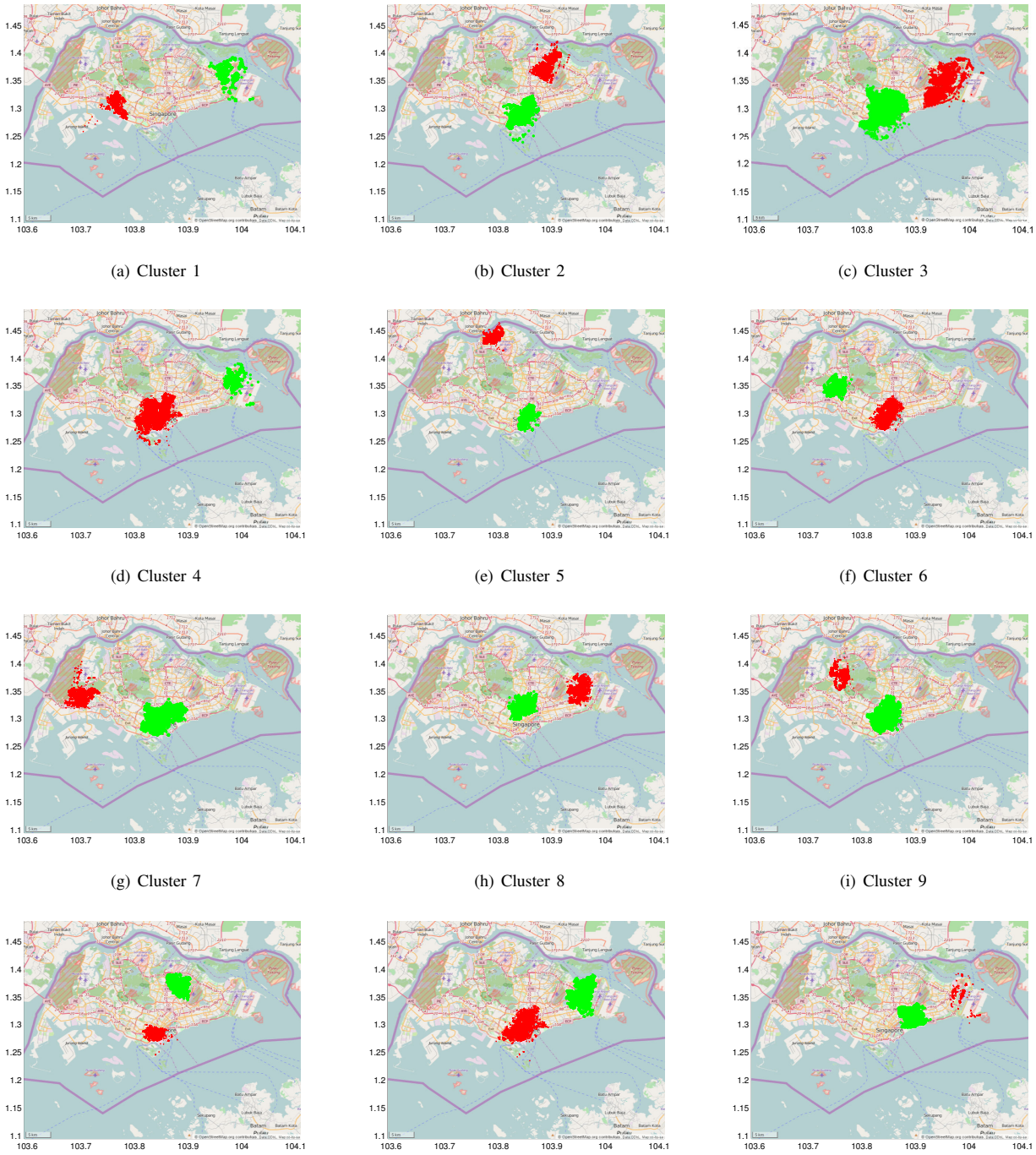
$$D_{ij} = Distance(\{Lat_{o_i}, Lon_{o_i}\}, \{Lat_{o_j}, Lon_{o_j}\}) + Distance(\{Lat_{d_i}, Lon_{d_i}\}, \{Lat_{d_j}, Lon_{d_j}\}), \quad (1)$$

where $Distance(\{Lat_{o_i}, Lon_{o_i}\}, \{Lat_{o_j}, Lon_{o_j}\})$ is the distance (in km) between the two GPS points $\{Lat_{o_i}, Lon_{o_i}\}$ and $\{Lat_{o_j}, Lon_{o_j}\}$.

We perform the clusiVAT sampling step on the dataset X consisting of $N = 10,000,000$ trips using $k' = 100$ and $n = 2,000$, which gives us $2,000$ distinct trips which retain the approximate geometry of entire dataset in a 4-dimensional space. In the next step we use DBSCAN as the clustering algorithm using the following parameters: $\epsilon = 2$ and $MinPts = 7$. The DBSCAN parameters were chosen so that the cluster have reasonable spatial spread so that they represent a specific locality as origin and destination location and the number of trips in each luster is also quiet high. After this step the trips which were not chosen in the clusiVAT sampling step are assigned to the cluster in which their nearest sampled trip belongs. Figs. 2 and 3 shows the 24 clusters obtained using this procedure. The green dots represent the origin and the red dots represent the destination of the trip.

From the discovered taxi trip clusters we can see that most of them either start or end in the center south area of Singapore. This is in accord with common sense, because that area is the *Central Business District* (CBD), and a lot of citizens need to travel to their offices in the morning and back from there in the evening. It is clearly seen that the counterpart of the city center in those clusters are all the dense residential areas.

In addition, the clusters also disclose some travel patterns that do not involve the city center. For example, Cluster 1 shows that many travelers travel from Tampines to Clementi, both of which are residential areas. This insight may reveal that the association, in terms of crowd movement between these two areas is stronger than other residential area pairs. Cluster 12 and 19 shows that many citizens travel from Geylang, which is a famous place for local dining and entertainment, to Tampines and Clementi, but not the other way around. That perhaps because people normally go to Geylang by public transport, e.g., bus or metro, and prefer taking a taxi home after dining or entertaining. Cluster 20 shows the crowd movement from Ang Mo Kio to Tampines. This is probably because there



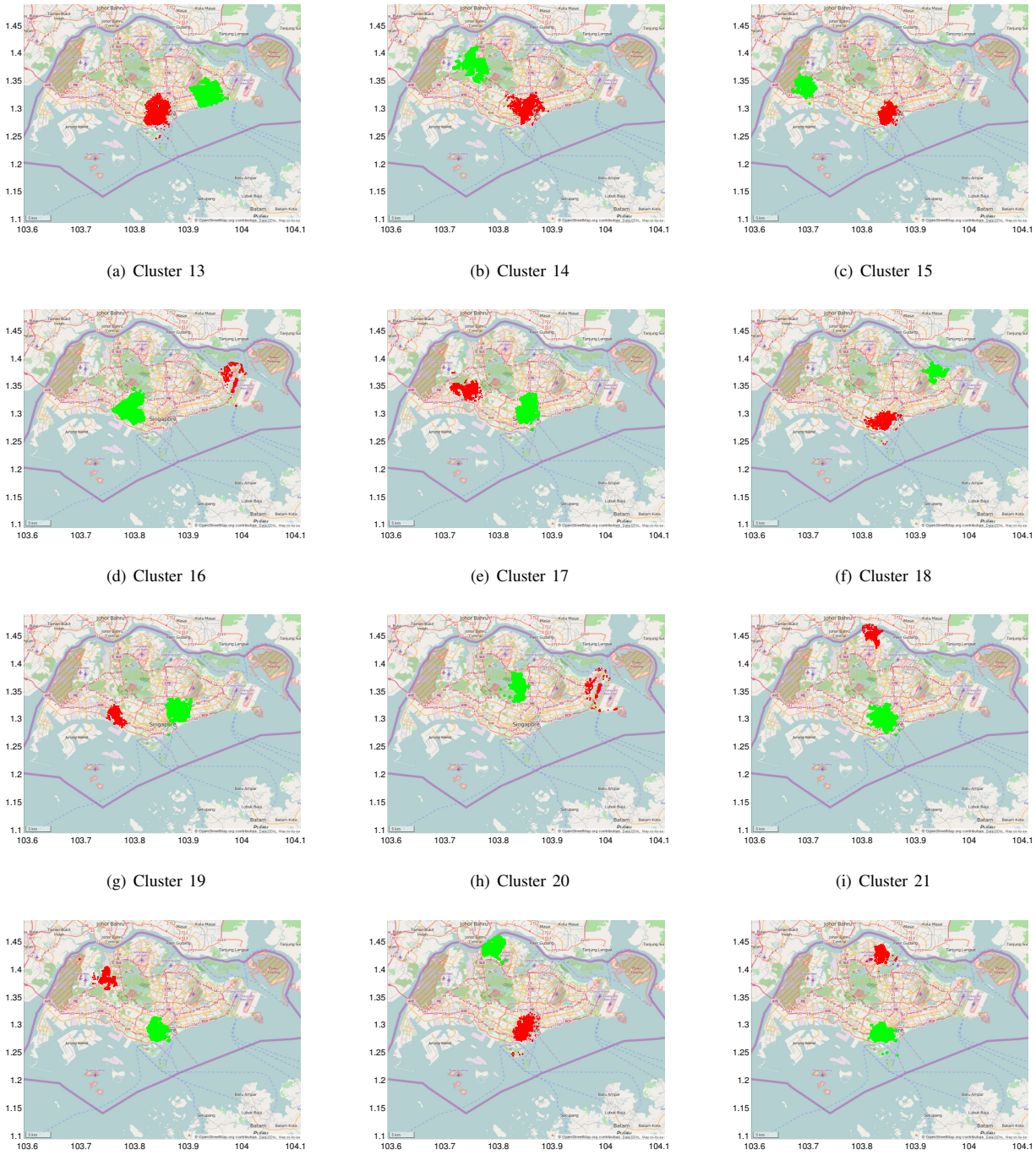
(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

(j) Cluster 10 (k) Cluster 11 (l) Cluster 12

Fig. 2. Clusters 1-12 obtained using DBSCAN (Green dots represent trip origin and red dots represents destination)



(a) Cluster 13 (b) Cluster 14 (c) Cluster 15

(d) Cluster 16 (e) Cluster 17 (f) Cluster 18

(g) Cluster 19 (h) Cluster 20 (i) Cluster 21

(j) Cluster 22 (k) Cluster 23 (l) Cluster 24

Fig. 3. Clusters 13-24 obtained using DBSCAN (Green dots represent trip origin and red dots represents destination)

is no convenient public transport from these two areas (the existing public transport to connect these two places requires a detour via the city center), and thus many citizens choose to use taxi.

Overall, the clustering result is validated from our background knowledge of the Singapore city, and further reveal some non-obvious insights of the crowd movement (e.g. Cluster 1, 12, 19 and 20). By tuning the parameters, we may find other clusters which can help to understand the urban mobility further. This will be our future work.

VI. Conclusions

In this paper we performed clustering of origin-destination GPS location of passenger taxi trips of the large scale Singapore taxi data, which consists of the trajectories of $15,000$ taxis collected over a duration of 1 month and have more than 370 million datapoints. From this dataset we extract more than 10 million origin-destination taxi trips on which clustering experiment is performed to understand urban mobility patterns. We refrain from using the centroid based clustering schemes as the information about the number of clusters, k and the general shape of clusters is not known. Since the dataset is large (more than 10 million trips), we can not apply hierarchical or density based clustering schemes directly as they are computationally prohibitive. To solve this problem we use the clusiVAT sampling scheme proposed in [2]–[4], which extracts the coarse cluster structure from the dataset which describe the major crowd movement and samples those datapoints which truly represent the structure of the dataset. We next use DBSCAN clustering algorithm on the sampled trips to extract the clusters and assigned the non sampled trips of the big dataset to the nearest cluster.

Using the above described procedure we were able to extract 24 trip clusters from the dataset which are compact enough to draw meaningful conclusions about the city mobility patterns, urban hot-spots and road network usage. The number of trips in each cluster is also large enough to be representative of the general traffic movement. We validate the clustering results from our background knowledge of the Singapore city and further reveal some non-obvious insights of the crowd movement. In the future we would experiment on fine-tuning the parameters to obtain clusters which can help to discover more insights into urban mobility.

VII. Acknowledgement

We thank the support from EU FP7 SocIoTal and H2020-ICT-2014-1 OrganiCity.

References

[1] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31(3), pp. 264–323, 1999.

[2] R. Hathaway, J. Bezdek, and J. Huband, "Scalable visual assessment of cluster tendency for large data sets," *Pattern Recognition*, vol. 39, pp. 1315–1324, 2006.

[3] D. Kumar, M. Palaniswami, S. Rajasegarar, C. Leckie, J. Bezdek, and T. Havens, "clusiVAT: A mixed visual/numerical clustering algorithm for big data," in *IEEE International Conference on Big Data*, Oct 2013, pp. 112–117.

[4] D. Kumar, J. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. Havens, "A hybrid approach to clustering in big data," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–1, 2015.

[5] M. Ester, H. peter Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996, pp. 226–231.

[6] K. Yamamoto, K. Uesugi, and T. Watanabe, "Adaptive routing of cruising taxis by mutual exchange of pathways," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer Berlin Heidelberg, 2008, vol. 5178, pp. 559–566.

[7] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti, "Taxi-aware map: Identifying and predicting vacant taxis in the city," in *Ambient Intelligence*. Springer Berlin Heidelberg, 2010, vol. 6439, pp. 86–95.

[8] H.-w. Chang, Y.-c. Tai, and J. Y.-j. Hsu, "Context aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Min.*, vol. 5, no. 1, pp. 3–18, Dec. 2010.

[9] X. Wan, J. Wang, Y. Du, and Y. Zhong, "Dbh-clus: A hierarchal clustering method to identify pick-up/drop-off hotspots," in *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, May 2015, pp. 890–897.

[10] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering spatial patterns in origin-destination mobility data," *Transactions in GIS*, vol. 16(3), pp. 411–429, 2012.

[11] X. Zhu and D. Guo, "Mapping large spatial flow data with hierarchical clustering," *Transactions in GIS*, vol. 18, no. 3, pp. 421–435, 2014. [Online]. Available: http://dx.doi.org/10.1111/tgis.12100

[12] F. Mao, M. Ji, and T. Liu, "Mining spatiotemporal patterns of urban dwellers from taxi trajectory data," *Frontiers of Earth Science*, pp. 1–17, 2015.

[13] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces," in *ACM Conference on Ubiquitous Computing*. New York, NY, USA: ACM, 2012, pp. 943–949.

[14] X. Wan, M. Gao, J. Kang, and J. Zhao, "Taxi origin-destination areas of interest discovering based on functional region division," in *International Conference on Innovative Computing Technology (INTECH)*, Aug 2013, pp. 365–370.

[15] R. Sibson, "Slink: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16(1), pp. 30–34, 1973.

[16] D. Wishart, "Mode Analysis: A Generalization of Nearest Neighbor which Reduces Chaining Effects," in *Numerical Taxonomy*, 1969, pp. 282–311.

[17] G. Ball and D. Hall, "ISODATA, a novel method of data anlysis and pattern classification," *Tech. rept. NTIS AD 699616. Stanford Research Institute, Stanford, CA*, 1965.

[18] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, p. 129–137, 1982.

[19] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999. [Online]. Available: http://doi.acm.org/10.1145/304181.304187

[20] J. Bezdek and R. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," *International Joint Conference on Neural Networks (IJCNN)*, vol. 3, pp. 2225–2230, 2002.