A New Hierarchical Clustering Approach For Sparse Mobile Phone Trajectories

Weixi Wang [1,2], Zhaoliang Luan[2, 4], Biao He[1, 2, *], Xiaoming Li[1,2], Dejing Zhang [5], Zhengdong Huang [2,3], Wei Tu[1, 2, 3]

[1] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518040, China - measurer@163.com, whu_hebiao@hotmail.com, lxmingster@163.com

[2] Shenzhen Key Laboratory of Spatial Information Smart Sensing and Services and Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China - (zdhuang, tuwei)@szu.edu.cn

[3] Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and GeoInformation, Shenzhen University, Shenzhen 518060, China - tuwei@szu.edu.cn

[4] College of Civil Engineering, Shenzhen University, Shenzhen 518060, China - luan_zhaoliang@163.com

[5] School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, China - djzhang@whu.edu.cn

Commission VI, WG IV/10

KEY WORDS: Trajectory, Mobile phone data, Human activities, Hierarchical clustering, Human behavior

ABSTRACT:

Understanding the pattern of human activities benefits both the living service providing for the public and the policy-making for urban planners. The development of location-aware technology enables us to acquire large volume individual trajectories with different spatial and temporal resolution, such as GPS trajectories, mobile phone positioning data, social media check-in data, Wifi, and Bluetooth. However, the highest population penetrated mobile phone positioning trajectories are hard to infer human activity pattern directly, because of the sparsity in both space and time. This article presents a hierarchical clustering approach by using the move and stay sequences inferred from spare mobile phone trajectories to uncover the hidden human activity pattern. Personal stays at some places and following moves are first extracted from mobile phone trajectories, considering the spatial uncertainty of position. The similarity of trajectories is measured with a new indicator defined by the area of a spatial-temporal polygon bound with normalized trajectories. Finally, a hierarchical clustering method is developed to group trajectories with similar stay-move chains from the bottom to the top. The obtained clusters are analyzed to identify human activity patterns. An experiment with mobile phone users’ one-day trajectories in Shenzhen, China was conducted to test the performance of the proposed clustering approach. The results indicate all used trajectories are classified into 10 clusters representing typical daily activity patterns from the simple home-staying mode to complex home-working-social activity daily cycles. This study not only unravels the hidden activity patterns behind massive sparse trajectories but also deepens our understanding of the interaction of human activity and urban space.

1. INTRODUCTION

Human activities are complex because of the diversity of personal social-economical characteristics and the heterogeneity of geographical environment. Understanding human activity pattern draws a meaningful portrait of the movements of humans and their interactions with the geographical space (Sagl et al., 2014). It benefits both the daily service providing for the public and the policy-making for urban planners. Traditional human activity pattern related studies heavily rely on the reliable travel survey (Axhausen and Gärling, 1992; Shen et al., 2014). Basically, travel survey investigates a set of volunteers to fill a form recording their travel, including start time, departure location, travel route, destination, arriving time, stay places, and duration of stay. But travel survey is labor-intensive, time-consuming, and low penetrated. For example, travel survey every five years in the city of Shenzhen, China will cost more than 2 months to find the volunteers and clean survey forms. Automatic collecting human activities information and unraveling the hidden human activity pattern is always an important topic in transportation and urban science (Yuan et al., 2017).

The rise of location-aware technologies, i.e., GPS, Wifi, Bluetooth, etc., has enabled us to track human in both outdoor and indoor environments (Tu et al., 2010; Zhou et al., 2017; Yue et al., 2014). Therefore, a large volume of human trajectories with different spatial-temporal resolutions have been collected, such as taxi GPS points, smart card data, mobile phone positioning data, and social media check-in data. These trajectories have provided unprecedented opportunities for big data-driven urban studies, for example, inferring the urban functions from mobile phone data (Tu et al., 2017, 2018), visualizing the urban dynamic landscape (Carlo, 2006), mapping the population (Deville et al., 2014), estimating the traffic congestion on the road (Castro et al., 2012), and etc. With respect to human activities, trajectories contain both position and time of an individual; therefore, they imply daily human activities, including home-in, working, education, shopping, etc. However, it is not easy to uncover the pattern from massive human trajectories. There are three reasons for this: 1) human activities are not explicitly described in the raw trajectory as there is no semantic label to assistant activity pattern discovery; 2) people are of different social-economical characteristics and their daily activities are constrained by heterogenous geographical environment, thereby human activity may be different in different urban areas; 3) trajectories may be very sparse during the collection phase. Taking mobile phone call detail record (CDR) trajectory generated by cellular network location technology as an example, the spatial accuracy may be

* Corresponding author

up to 500 meters because of the cellular network positioning, while the temporal accuracy may be more than several hours.

Clustering groups a set of objects in such a way that objects within the same group are more like each other than to those in other groups. It is an effective approach to reveal hidden knowledge from massive data without prior information (Iovan et al., 2013; Jain and Dubes, 1998). Basically, trajectories clustering contains two main parts: the measure of trajectories similarity and the clustering method. The trajectory similarity quantifies how much two trajectories are the same. The Euclidean distance, the temporal distance, and the semantic distance measure the similarity from the spatial, temporal, and semantic view respectively, if trajectories contain semantic labels. The edit distance is defined as the minimum number of operations needed to transform one trajectory to another (Yuan et al., 2014). The dynamic time wrap distance (Berndt and Clifford, 1994) is a measure that searches the optimal warping path between two series by adjusting the alignment. The longest common subsequence (LCSS) (Vlachos et al., 2002) use the most common sub trajectory to indicate the similarity. The weighted spatial-temporal-semantic distance (Wan et al., 2012) weights the spatial, temporal, and semantic distance to quantify the similarity of two trajectories.

The clustering methods contain four categories, including connectivity-based, centroid-based, distribution-based, density-based clustering approaches, etc (Lee et al., 2007; Lin and Hsu, 2014; Xie and Zhao, 2017). The connectivity-based clustering groups from the bottom to the top such that objects are more related to nearby objects than to objects farther away. In the centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. The similarity between objects is defined as an object to the central vector, rather than original objects. The distribution-based clustering groups an object to the cluster with the most similar distribution. In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in low-density areas are usually considered to be noise and border points. These methods have achieved success on the high-resolution GPS trajectories but leave a room to be improved on sparse mobile phone trajectories with a higher spatial-temporal uncertainty.

Mobile phone positioning data is generated when a mobile phone user makes a call or accesses to the Internet. Because of the practical cellular network positioning technology, passive mobile phone data is spatial sparse (Jiang et al., 2013). The uncertainty of mobile phone location is large, about 100 meters in the high density-built area but 500 meters in low-density rural area. On the other hand, mobile phone users with similar daily activities have quite different mobile phone positioning trajectories for their different geographical environments. To overcome the inherent sparsity of mobile phone trajectories, a hierarchical clustering approach is proposed to group trajectories into the same clusters. Stops and moves are firstly identified to smooth the uncertainty of location. A new indicator is proposed to measure the similarity of two trajectories with help of the bounded spatial-temporal polygon. The hierarchical clustering is used to group mobile phone trajectories with similar daily activity sequences. Finally, the frequent activity patterns behind large volume massive mobile phone trajectories are revealed. An experiment using more than 1 million users' mobile phone trajectories in Shenzhen, China was conducted to evaluate the performance of the proposed approach. The results indicate all used trajectories are classified into 10 clusters representing typical daily activity patterns from the simple home-staying mode to complex home-working-social activity daily cycles. This study not only unravels the hidden activity patterns behind massive sparse trajectories but also deepens our understanding of the interaction of human activity and urban space.

The remainder of this article is as follow. Section 2 introduces the methodology, including the stop and move detection, the trajectory similarity measure, and the hierarchical clustering method. Section 3 describes the experiment and the results. Section 4 concludes this article.

2. METHODOLOGY

The proposed approach is to group sparse mobile phone trajectories by considering the spatial-temporal sparsity of trajectories, instead of a direct clustering based on the geometry shape. It contains three main steps, including move and stays activity detection, trajectory similarity and hierarchical clustering. Figure. 1 shows the workflow of the presented approach. Virtual human activities including moves and stays are first detected by considering the uncertainty of location. The similarity of mobile phone trajectories is defined with the spatio-temporal polygons. The hierarchical clustering is used to group mobile phone trajectories with similar daily activity sequences. Consequently, the hidden human activity patterns are revealed.

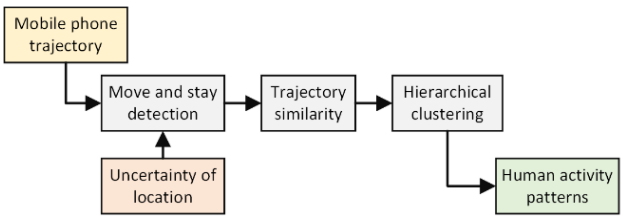

Figure 1. The workflow to clustering mobile phone positioning trajectory

2.1 Move and stay detection

Move and stay detection extracts reliable human stay and move from sparse mobile phone positioning data to overcome the positioning error, about 100~500 meters. Mobile phone positioning records of an individual are first sorted according to the recording time. Then, following the time stamps, mobile phone positioning records of a user are connected as a spatial-temporal trajectory. Then, the trajectory is segmented by the positions of two consecutive points. Basically, if two consecutive records are at the same location, in other words, the person does not move, a potential activity is found, such as p1–p2, p6–p7, p8–p9 in Figure 2b.

Because of the low positioning accuracy and frequency of cellular network based positioning technology, mobile phone trajectories are spare in both space and time. On the other hand, consecutive points of mobile phone positioning data will jump between adjacent cell stations, even though the user stays at the same place. Therefore, false potential human activities may be inferred. To overcome this issue, a distance threshold d is used to filter false moves: if the distance from the current point p to the previous mobile phone positioning point q is less than a given threshold d, the move may be false detected, and the current point can be merged into previous potential stay point. After sequentially processing all mobile phone positioning points, potential stay without type information and the associated move is revealed from raw trajectories.

It should be noted here that, because of the low positioning frequency, the interval between two consecutive mobile phone positioning records may be very long, which will significantly overestimate the travel time. For example, if there is one user who makes only one phone call at his or her workplace, the duration of the corresponding activity may be 0. Figure 3 gives an example of this case. The point p_1 is recorded at time 14:00:00, while previous point p_0 and the following p_2 are recorded at 8:00:00 and 20:00:00 respectively.

To overcome this low sampling character of mobile phone positioning trajectory, we use the high frequently visited places to infer the possible stay. We firstly summarize the frequency of locations and identify the first *kth*-location as important places, i.e., home, workplace, shopping stores, etc. Considering the rhythm of human daily activities, if there is one point in these places, we still recognize it as a stay, rather than an intermediate point travel. To adjust the start time at a possible stay point p_i, we estimate the travel time from previous GPS point p_{i-1} to this location p_i as equation (1) using the average travel speed v in the city. Then, the adjusted start time of a move from p_{i-1} to p_i is as equation (2). The adjusted end time of this move at location p_i is defined by equation (3). Figure 3 demonstrates the adjustment of travel. Virtual mobile phone records $q_0 - q_3$ are inserted into the raw trajectory to alleviate the low sampling shortcoming. q_0 and q_1 are inserted between p_0 and p_1, while q_2 and q_3 are inserted between p_1 and p_2. The location of q_0 is the same with p_0. The location of q_1 and q_2 are the same as the target point p_1. The location of q_3 is the same as the following point p_2. The time of $q_0 - q_3$ is calculated by equation (2) and (3) respectively.

$$t = \text{distance}(p_{i-1}, p_i)/v \quad (1)$$

Where p_i is the stay point,
p_{i-1} is previous GPS point,
distance is the Euclidian distance from p_{i-1} to p_i,
v is the average travel distance in the city.

$$t_{start} = (t_{pi} + t_{pi-1})/2 - t/2 \quad (2)$$

Where t_{pi} is the time at point p_i,
t_{pi-1} is the time at point p_{i-1},
t is travel time from p_i to p_{i-1}.

$$t_{end} = (t_{pi} + t_{pi-1})/2 + t/2 \quad (3)$$

Where t_{pi} is the time at point p_i,
t_{pi-1} is the time at point p_{i-1},
t is travel time from p_i to p_{i-1}.

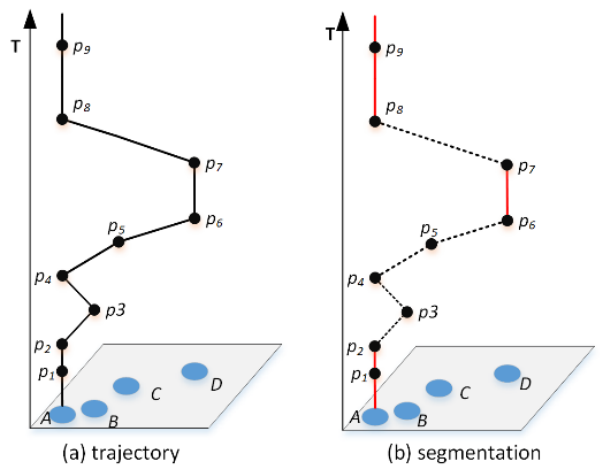

Figure 2. The detection of move and stay

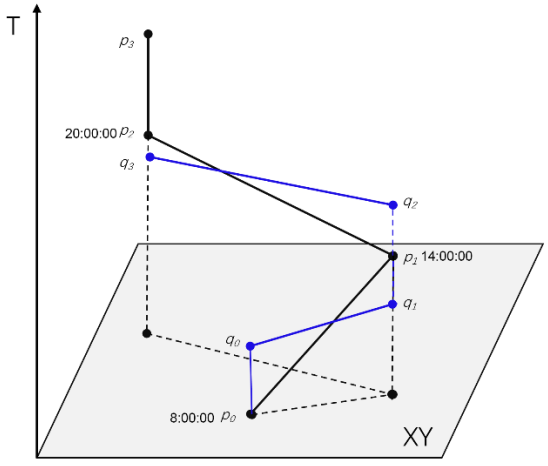

Figure 3. The adjust of move between stops

2.2 Trajectory similarity

Considering the sparsity of mobile phone trajectories, the spatial-temporal similarity indicates the similarity of the adjusted trajectories, rather than raw trajectories. It calculates the spatio-temporal polygon area bounded by the two adjusted trajectories, as equation (4). Firstly, mobile phone positioning points are normalized using the maximum travel radius of the mobile phone user. Then, virtual points are linearly interpolated with the same sampling intervals. The obtained sequential points are used to assist the calculation. Figure 4 gives an example to illustrate the similarity measure. It can be seen the more similarity between two trajectories, the less the area. If two trajectories are the same, the similarity is equal to zero.

$$S = \int_{i=1}^{N} \Delta t \cdot \left(d_{i,j} + d_{i+1,j+1}\right)/2.0 \quad (4)$$

Where Δt denotes the time step,
N denotes the number of the time step,
d_{ij} denotes the distance from i to j.

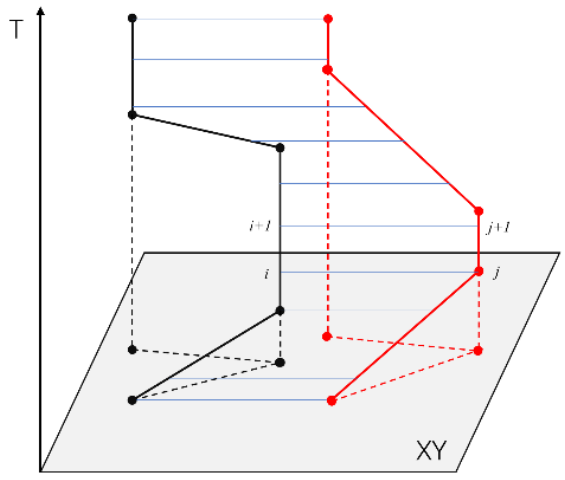

Figure 4. The similarity of trajectories

2.3 The hierarchical clustering

The hierarchical cluster method was used to group trajectories from the bottom to the top to uncover the hidden human activity patterns. The ward clustering (Ward, 1963) minimizes the increased distance of within-cluster distance by merging two clusters. The within-cluster distance is defined as equation (5),

which is the sum of the similarity of a trajectory to the cluster center.

$$D_A = \sum_{i \in A}^{m} \|i - \bar{m}_A\|^2 \quad (5)$$

Where i is a trajectory in cluster A,
$\bar{m}_A$ is the center of cluster A,
m is the number of elements.
$\|\cdot\|$ denotes the similarity between two trajectories.

The increased distance of merging two clusters is defined as equation (6):

$$\sum_{i \in A \cup B}^{m+n} \|i - \bar{m}_{A \cup B}\|^2 - \sum_{i \in A}^{m} \|i - \bar{m}_A\|^2 - \sum_{i \in B}^{n} \|i - \bar{m}_B\|^2 \quad (6)$$

Where i is a trajectory in cluster A or cluster B,
$\bar{m}_A$ and $\bar{m}_B$ are the centres of clusters,
m and n are the numbers of elements in cluster A and B.
$\|\cdot\|$ denotes the similarity between two trajectories.

By grouping all trajectories from the bottom to the top, a set of clusters are obtained. The patterns of human activities are analyzed with the center of trajectories in the same cluster. We plot the cluster center to extract daily activity patterns.

3. EXPERIMENT AND ANALYSIS

The experiment was conducted in Shenzhen, China to evaluate the performance of the proposed hierarchical clustering approach. Shenzhen is a modern metropolitan with 15 million population, covering 1996 square km. The used mobile phone positioning dataset was provided by a dominated mobile communication company. The locations of mobile phone users were recorded when they made phone calls or send messages. Every trajectory has at least 6 points in a day. More than 100 thousand trajectories were grouped using the presented hierarchical clustering approach.

The obtained clusters are displayed in Figure 5. It demonstrates that these trajectories were grouped into 10 main clusters with one to more than 5 visiting places. The first cluster has only one place. It indicates that the user does not travel in the whole day, just stay in one place. The corresponding users may be children or old adults spending most time at home. Another possible user group for this is the people working at home, who don't travel for the whole day. The cluster 2 contains two places suggesting the corresponding user travels routinely between their home and workplace.

The remained clusters have more than 2 visiting places. Cluster 3-5 are with three visiting places, which indicate the users travel between home, workplaces, and one non-home and non-work places (e.g. shopping center, supermarkets, etc.) in the day. These users visit the third places after working. Cluster 6 to 8 have 4 visiting places, with two non-home and non-work places. They demonstrate that their users have diverse activities in a day. The most complex activity patterns are displayed in cluster 9 and 10, with more than 4 visiting places. They suggest the corresponding users have a complex daily activity and mobility network.

We further summarized the count of trajectories in each cluster. Table 1 reports the summary information. It can be seen most of the mobile phone user (40.3%) are grouped into cluster 2 with a simple home-work activity pattern. The least users (0.16%) belong to cluster 8 with four visiting places with a circular activity-mobility network. 25.5% of users stay at one places without any travel. Many users (91.27%) visit less than 4 places in a day, just travel between home, work, and a non-home non-work place, as the cluster 1-5 show.

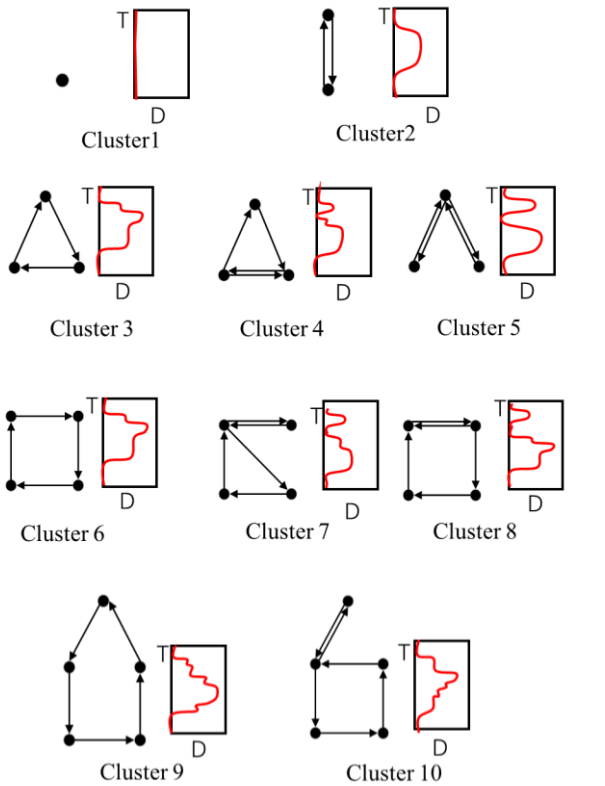

Figure 5. Trajectory clusters. A node denotes one visiting place. An edge denotes a move between places. T is the time dimension. D is the distance from the centre of a user' trajectory

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Ratio	25.5	40.3	18.3	5.89	1.28
	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Ratio	5.4	1.92	0.16	1.0	0.25

Table 1. The ratio of trajectory clusters

4. CONCLUSION

Understanding the pattern of human activities benefits both the living service providing for the public and the policy-making for urban planners. The development of location-aware technology enables us to acquire large volume individual trajectories with different spatial and temporal resolution. However, the highest population penetrated mobile phone positioning trajectories are hard to infer human activity pattern directly, because of the sparsity in both space and time. Considering the spatial-temporal sparsity of mobile phone trajectory, this article presents a hierarchical clustering approach through virtual human activity to uncover the activity pattern. Personal stays at some places and following moves are first extracted from mobile phone trajectories, considering the spatial uncertainty of position. The similarity of trajectories is measured with a new indicator defined by the area of a spatial-temporal polygon bound with normalized trajectories. Finally, a hierarchical clustering method is developed to group trajectories with similar stay-move chains from the bottom to the top. The obtained clusters are analyzed to identify human activity patterns.

An experiment with large volume of mobile phone users’ one-day trajectory in Shenzhen, China was conducted to test the performance of the proposed approach. The results indicate all trajectories are classified into 10 clusters representing typical daily activity patterns from home-staying to complex home-working-social activity mode. This study not only uncovers the hidden activity patterns behind massive trajectories. This obtained results also deepen us understand the using of geographical space. They demonstrate that most people stay at home, go to work, possibly with one non-home and no-work activity. Therefore, for innovative applications in urban planning or smart cities, we should pay much attention to the inherent human activity patterns.

ACKNOWLEDGMENTS

This research was supported in part by Natural Science Foundation of China (No. 41401444), the open fund of the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, China (KF-2016-02-010), and Shenzhen Scientific Research and Development Funding Program (#JCYJ20170412105839839, #CXZZ201505041416 23042).

REFERENCES

Sagl, G., Delmelle, E., and Delmelle, E. Mapping collective human activity in an urban environment based on mobile phone data. *Cartography & Geographic Information Science*, 2014, 41(3): 272-285.

Axhausen, K. W., and Gärling, T. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews*, 1992, 12, 323-341.

Shen, L., and Stopher, P. R. Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 2014, 34, 316-334.

Yuan, G., Sun, P., Zhao, J., Li, D., Wang, C. A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, 2017, 47, 123-44.

Yue, Y., Lan, T., Yeh, A.G.O., Li, Q.Q. Zooming into individuals to understand the collective: a review of trajectory-based travel behavior studies. *Travel Behaviour and Society*, 2013, 1(2), 719–723.

Zhou, B.; Li, Q.; Mao, Q.; Tu, W. A Robust Crowdsourcing-Based Indoor Localization System. *Sensors*, 2017, *17*, 864.

Tu W, Fang, Z. LI, Q.Q. Exploring time-varying shortest path of urban OD pairs based on floating car data. *18th International Conference on Geoinformatics, Geoinformatics 2010*. 2010. Beijing, China.

Tu, W., Cao, J., Yue, Y., Shaw S.L., Zhou, M., Wang Z.S., Chang X.M., XU, Y., LI, Q.Q. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*. 2017, 31(12), 2331-2358.

Tu W., Hu, Z.W., Li, L.F., Cao, J.Z., Jiang, J.C, Li, Q.P, Li, Q.Q. Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data. *Remote sensing*. 2018, 10(1), 141.

Ratti, Carlo. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*. 2006, 33(5): 727-748.

Tu, W, Cao, R., Yue, Y., Zhou, B., Li, Q.P, Li, Q.Q. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *Journal of Transport Geography*. 2018, 69, 45-57.

Deville, P., Linard, C., Martin, S. Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., and Tatemet, A.J. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 2014, 111, 15888-15893.

Castro, P. S., Zhang, D., and S. Li. Urban traffic modeling and prediction using large-scale taxi GPS traces. in *International Conference on Pervasive Computing*, 2012, pp. 57-72.

Iovan C, Olteanu-Raimond A M, Couronné T, et al. Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. *Lecture Notes in Geoinformation & Cartography*, 2013: 247-265.

Jiang S, Fiore G A, Yang Y, et al. A review of urban computing for mobile phone traces:current methods, challenges and opportunities[C]// *ACM SIGKDD International Workshop on Urban Computing*. ACM, 2013:1-9.

Yuan Y, Raubal M. Measuring similarity of mobile phone user trajectories – a Spatio-temporal Edit Distance method. *International Journal of Geographical Information Science*, 2014, 28(3): 496-520.

Jain, A. K., and Dubes, R. C. Algorithms for clustering data. 1988.

Berndt, D. J., and Clifford, J. Using dynamic time warping to find patterns in time series. in KDD workshop, 1994, pp. 359-370.

Vlachos, M., Kollios, G., and Gunopulos, D. Discovering similar multidimensional trajectories. *Proceedings of 18th International Conference on in Data Engineering*. 2002, pp. 673-684.

Wan, Y., Zhou, C., Pei, T. Semantic-Geographic Trajectory Pattern Mining Based on a New Similarity Measurement. *ISPRS Int. J. Geo-Inf.* 2017, *6*, 212.

Lee, J.-G., Han, J., and Whang, K.-Y. Trajectory clustering: a partition-and-group framework. *Proceedings of the 2007 ACM SIGMOD international conference on Management of Data*, 2007, pp. 593-604.

Lin, M., and Hsu, W.-J. Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing*, 2014, 12, 1-16.

Xie, M., and Zhao, Q. TaxiCluster: A Visualization Platform on Clustering Algorithms for Taxi Trajectories. *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*. 2017, pp. 138-147.

Ward, J. H., Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 1963, 58, 236–244.