A framework for spatio-temporal clustering from mobile phone data

Yihong Yuan[a,b]

[a]Department of Geography, University of California, Santa Barbara, CA, 93106, USA
yuan@geog.ucsb.edu

Martin Raubal[b]

[b]Institute of Cartography and Geoinformation, ETH Zurich, 8093 Zurich, Switzerland
mraubal@ethz.ch

1 Introduction

When analyzing quantitative spatial data, it is often essential to classify spatial objects into sub-groups, so that objects within the same group are more similar to each other than those in different groups. One example is the partitioning of a study region into 'poor' versus 'rich' according to the average income [1]. Moreover, since the temporal dimension is an important factor for most social activities, researchers have developed various spatio-temporal clustering techniques for classifying observations that show similar behavior in both spatial and temporal dimensions. These techniques have been applied in many application fields, such as trajectory clustering [2], crime analysis [3] and epidemiology modeling [4].

The recent rapid development of Information and Communication Technologies (ICTs) has created a wide range of novel spatio-temporal data sources (e.g., georeferenced mobile phone records) for researchers to explore the travel behavior and dynamic mobility patterns of phone users [5, 6]. These can be used as informants of human activity including long term choices such as where to live (work and go to school) and shorter term choices such as activity scheduling in a week or even daily. Generally, mobile phones are capable of recording location information by several ways such as using Global Positioning System (GPS), service-provider assisted faux GPS or simply by logging the connected cellular tower information. In the last case, a mobile phone has to emit signals for contacting a nearby cell phone tower in order to be located [7]. Tracking information is then only available when a phone call occurs and the spatio-temporal points are under relatively low temporal resolution and unequally spaced on the time axis. This falls into the category of "event-based positioning techniques" as discussed in [8]. Since the collected data are normally generated as scattered sample points, further analysis is required to identify the inherent mobility patterns. Applying clustering techniques to these datasets can facilitate the extraction of the spatio-temporal characteristics of user mobility patterns. In this paper, we investigate four categories of spatio-temporal clustering methods that can be applied to mobile phone datasets at various spatio-temporal scales. The results can be utilized as a reference framework and theoretical basis for clustering human mobility patterns and for conducting spatio-temporal data mining in the age of instant access.

2 Spatio-temporal clustering based on mobile phone data

Mobile phones and other wireless devices collect large numbers of measurements about their users. Although the completeness and accuracy vary for different datasets, a typical mobile phone dataset contains the following three categories of user information: (1) Spatio-temporal tracking information (i.e., the time and approximated location of phone calls); (2) service usage information (i.e., the frequency and duration of voice, text, and other types of service usage); and (3) demographic profiles if available (including individual-level profiles, such as age or gender, and super-individual-level profiles, such as social conditions or cultural backgrounds) [9].

For spatio-temporal clustering, researchers mostly focus on the 1st type of information (spatio-temporal tracking information). However, the 2nd and the 3rd categories can also be utilized to enrich the background information for clustering analysis. For example, Yuan et al. [10] utilized a dataset from northeast China, which covers over one million people and includes mobile phone connection records for a time span of 9 days. It includes the time, duration, and approximate location of mobile phone connections, as well as the age and gender attributes of the users. Table 1 provides several sample records. The phone number, longitudes and latitudes are not shown for reasons of privacy.

Table 1: Sample records from the example data set

Phone #	1350*******
Longitude	126.*****
Latitude	45.*****
Time	14:26:24
Duration	12mins
Receiver phone #	1360*******

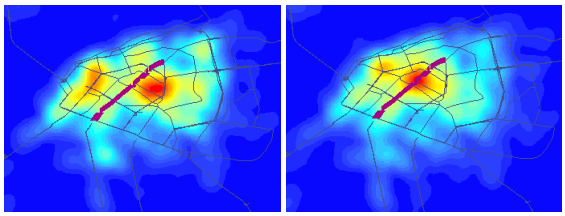
Phone #	Gender	Age
1350*******	Male	30

As a standard procedure of clustering analysis, spatio-temporal clustering should also be conducted based on the following three steps: 1) Select variables and features for clustering; 2) decide which clustering algorithms to employ; and 3) interpret clustering results [11]. Sections 2.1-2.4 discuss four types of spatio-temporal clustering under different spatial and temporal scales based on geo-referenced mobile phone datasets. The first two types of clustering are conducted on individual-level features (i.e., user trajectories), whereas the latter two types concentrate on clustering urban-level features (i.e., hourly mobility counts in a certain area).

2.1 Intra-trajectory clustering

Intra-trajectory clustering refers to the clustering analysis conducted to derive the internal patterns of user trajectories. For instance, in Bagrow and Koren [12], the bimodal nature of human trajectories is investigated based on a large dataset of cellular telecommunication records. They quantified how much of a user's spatial dispersion is due to motion between clusters, and how clusters are spatially and temporally separated from one another.

Moreover, previous research has developed several methodologies for extracting '*stops*' (which are usually considered as the low-velocity parts) from moving object data (e.g., vehicle trajectories, animal tracking, etc.), and all of these methods can be utilized to analyze the internal patterns of spatio-temporal tracking information in mobile phone datasets—see the CB-SMoT method (Clustering-Based Stops and Moves) introduced in [13]. Phithakkitnukoon *et al.* [14] also applied a method to identify the *stops* of mobile phone carriers by regrouping a trajectory into sub-trajectories based on predefined spatial and temporal thresholds on consecutive points. Yuan et al. [10] applied the same method and studied the correlation between and location of "*stops*" and the usage of mobile phone data (Figure 1). Once the *stops* have been extracted, the home location of each user is estimated as the most frequent *stop* during the night hours and the work location as the most frequent stop during day hours on weekdays. Figure 1 visualizes the distribution of home and work locations of users in City A. These POIs can also be combined with our previous research on user trajectory patterns to further examine the determinants of an individual's activity space [15]. However, as argued before, since the spatial and temporal resolutions of mobile phone records are relatively low, both accuracy and precision need to be critically taken into account when conducting *stops* extraction from mobile phone data.



(a) (b)

Figure 1: Clustering of (a) home locations and (b) work locations based on mobile phone records [15]

In Moreno et al. [16], the researchers took one step further from the identification of stops. They developed an algorithm for analyzing the behavior of the moving object in order to infer the goal of the stop. All these methods can be very helpful for analyzing the internal patterns of spatio-temporal points generated from mobile phone datasets. For future research, an interesting topic would be analyzing how the distribution of *stops* correlates with the social attributes of phone users.

2.2 Inter-trajectory clustering

First we need to differentiate between two types of research: *moving clusters identification* and *trajectory clustering*. Moving clusters refer to a set of objects that move close to each other for a long time interval [17], while trajectory clustering focuses on classification and regrouping of multiple trajectories based on their shapes and other features. Objects in the same moving clusters usually have similar trajectories during the given time span; however, objects with similar trajectories do not necessarily need to be in the same moving clusters. In a majority of these studies unsupervised classification was employed as the method of analysis. Several algorithms, from basic clustering techniques such as k-means clustering and hierarchical clustering, to more advanced techniques such as Hidden Markov Model (HMM) and Principle Component Analysis (PCA), have been proposed to classify the motion patterns of objects in real-world applications, such as gesture recognition [18]. As discussed in previous research, the most important issue in clustering trajectory data is to identify the spatio-temporal attributes to be clustered on [8]. Skupin et al. [19] proposed a method to visualize and analyze space-time paths (SPA) in attribute space, which provides a novel perspective for modeling multidimensional attribute data. For mobile phone data, the classification of user trajectories can be helpful for understanding the movement patterns of different population groups; however, the complexity of the problem is exacerbated by the low resolution in both the spatial and temporal scales. Since imbedded GPS devices are only available in a small portion of cell phones (i.e., smart phones),

most mobile phone datasets only include the location of the base stations, and the tracking information is only available when a phone call occurs; therefore, the recorded spatio-temporal points have low temporal resolution and are unequally spaced on the time axis. This increases the complexity of clustering the trajectories of mobile phone users (Figure 2).

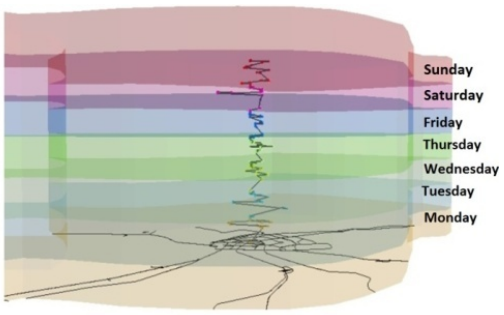


Figure 2: A week-long individual travel-activity path [9].

A potential research direction here is to explore how the uncertainty in both spatial and temporal scale affects the clustering analysis. For example, Nanni and Pedreschi [2] proposed a density based method that can be applied to cluster trajectories unevenly spaced in the spatial and temporal domain.

2.3 Intra-urban clustering

Since individuals are atoms in an urban system, the spatio-temporal characteristics of an urban system can be viewed as a generalization of individual behavior; therefore, mobile phone data also provide new insights into the analysis of the mobility patterns in urban systems. Researchers believe that urban structure has a strong impact on urban-scale mobility patterns, indicating that different areas inside a city are associated with different inhabitants' motion patterns [20]; hence, a potential research direction is to classify urban areas based on their dynamic mobility patterns. Unlike trajectory clustering discussed in sections 2.1 and 2.2, here the variables used for clustering are aggregated from individual trajectories in different regions. We only discuss clustering approaches that involve all three components: objects (i.e., phone users), space, and time.

Andrienko [8] et al. summarized the categories of spatial temporal aggregations (see Table 1 in [8]). Here we extend the table from a clustering perspective (see Table 2, T for time, S for space and A for attributes, i.e., the number of people at each location).

Table 2: Different scenarios and corresponding clustering features

Scenarios	Clustering feature
$S \rightarrow (T \rightarrow A)$ Time series of summary attribute values in each location	Time series
$T \rightarrow (S \rightarrow A)$ Summary attribute values associated with each time unit	Spatial series
$S \times S \rightarrow (T \rightarrow A)$ For each pair of locations, time series of flows between the locations by time intervals	Vector time series
$T \times T \rightarrow (S \rightarrow A)$ Each pair of time units, aggregate attributes representing changes between the spatial configurations	Spatial series of the change of aggregated attribute
$T \times T \rightarrow (S \times S \rightarrow A)$ For each pair of time units, aggregate moves (flows) of objects between locations	Matrix in spatial dimension

Various clustering techniques have been explored for each case listed in Table 2. For example, the hourly phone call frequencies can be viewed as regular time series. In this case many well-developed clustering techniques for time series data can be applied, such as Longest Common Subsequence (LCSS) described in [21]. However, there are still remaining issues to be studied, such as how to cluster the aggregate moves of objects in the last row of Table 2.

Although it is also feasible to conduct intra-trajectory and inter-trajectory clustering for users in different urban areas, this category of analysis mainly focuses on the clustering process conducted directly on the urban-scale features. It can also be applied for detecting abnormal mobility patterns in cities, as well as providing references for researchers and policy makers to investigate the functioning patterns of different urban areas.

2.4 Inter-urban clustering

Similar to intra-urban clustering described in Section 2.3, it is also feasible to classify multiple cities based on their dynamic mobility patterns. The classification of cities has been tackled from various perspectives: from functional to geometrical (i.e., size, shape, etc.) [22, 23]; however, there has not been an extensive study on classifying cities based on their internal spatio-temporal mobility patterns. The biggest challenge here is to select one or more representative variables for the mobility patterns of an entire city. Here we provide potential directions for future work.

One option is to aggregate all space-time points and generate a mobility time series for each city; however, this approach cannot represent the spatial heterogeneity inside urban areas. A potential solution is to divide each city into sub-regions so that the internal mobility heterogeneity can be preserved. Several research questions and issues are associated with this topic, including:

1) How to divide a city into sub-regions?

2) Since the objective is to classify cities based on their mobility patterns, how to eliminate / incorporate the effects of other variables when conducting the clustering analysis, such as:
    - Urban morphology variables (size, shape of cities, etc.);
    - Urban demographic variables (population, average income, etc.);
    - Mobility central tendency variables: Central tendency refers to a typical value of the distribution, such as the average movement radius for a given city and the major direction of mobility flows (e.g.; central to outbound or outbound to central).
3) What kinds of variables should be used to represent the spatio-temporal patterns of each sub-region? Table 2 provides a good start for this question.
4) How to integrate the patterns in sub-regions to represent the whole city?
5) What kinds of similarity measure and clustering techniques should be used?

3 Conclusions

In this paper we have summarized the categories of spatio-temporal clustering based on mobile phone datasets. Four types of clustering were discussed regarding their methodologies, issues, and future directions: intra-trajectory clustering, inter-trajectory clustering, intra-urban clustering, and inter-urban clustering. Generally, two major challenges are the selection of clustering techniques and the selection of clustering variables for each specific circumstance. A framework of potential research questions is provided as a guideline for our future research agenda on reducing the complexity of dynamic data and classifying similar patterns in mobile datasets.

References

[1] Miller, H. Geographic data mining and knowledge discovery: An overview, in *Geographic Data Mining and Knowledge Discovery (Second Edition)*, H.J. Miller and J. Han, Editors, CRC Press: London, pages 3-32, 2009.

[2] Nanni, M. and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*. 27(3): 267-289, 2006.

[3] Chandra, B., M. Gupta, and M.P. Gupta. A Multivariate Time Series Clustering Approach for Crime Trends Prediction. *2008 Ieee International Conference on Systems, Man and Cybernetics (Smc), Vols 1-6*: 891-895, 2008.

[4] Carpenter, T.E. Methods to investigate spatial and temporal clustering in veterinary epidemiology. *Preventive Veterinary Medicine*. 48(4): 303-320, 2001.

[5] Song, C.M., et al. Limits of predictability in human mobility. *Science*. 327(5968): 1018-1021, 2010.

[6] Miller, H.J. and J. Han. Geographic data mining and knowledge discovery. 2nd ed, Boca Raton, FL: CRC Press. 458 p., 2009.

[7] Brimicombe, A. and C. Li. Location-based services and geo-information engineering. Mastering GIS, Chichester, UK ; Hoboken, NJ: Wiley-Blackwell. xiv, 378 p., 2009.

[8] Andrienko, G., et al. A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages and Computing*. 22(3): 213-232, 2011.

[9] Yuan, Y. and M. Raubal. Spatio-temporal knowledge discovery from georeferenced mobile phone data. in *MPA'10 - 1st Workshop on Movement Pattern Analysis*. Zurich, Switzerland, 2010.

[10] Yuan, Y., M. Raubal, and Y. Liu. Correlating mobile phone usage and travel behavior - a case study of Harbin, China. *Computers, Environment and Urban Systems*. 36(2): 118-130, 2012.

[11] Rutkowski, L. Computational intelligence : methods and techniques. English ed, Berlin: Springer. xiii, 514 p., 2008.

[12] Bagrow, J.P. and T. Koren. Investigating Bimodal Clustering in Human Mobility. *International Conference on Computational Science and Engineering*. 4: 944-947, 2009.

[13] Palma, A.T., et al. A clustering based approach for discovering interesting places in trajectories. in *ACMSAC*. New York, NY: ACM Press, 2008.

[14] Phithakkitnukoon, S., et al., *Activity-aware map: Identifying human daily activity pattern using mobile phone data*, in *HBU 2010*, A.A. Salah, et al., Editors. 2010, LNCS, Springer: Heidelberg. p. 14-25.

[15] Yuan, Y. and M. Raubal. Extracting clustered urban mobility and activities from georeferenced mobile phone datasets. in *ISPRS Workshop on Spatio-Temporal Data Mining and Analysis (STDM'11)*. London, United Kingdom, 2011.

[16] Moreno, B., et al., *Looking Inside the Stops of Trajectories of Moving Objects*, in *GeoInfo 2010*. 2010. p. 9-20.

[17] Kalnis, P., N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. *Advances in Spatial and Temporal Databases, Proceedings*. 3633: 364-381, 2005.

[18] Lee, H.J., Y.J. Lee, and C.W. Lee. Gesture classification and recognition using principal component analysis and HMM. *Advances in Mutlimedia Information Processing - Pcm 2001, Proceedings*. 2195: 756-763, 2001.

[19] Skupin, A. Visualizing human movement in attribute space, in *Self-Organising Maps: Applications in Geographic Information Science*, P. Agarwal and A. Skupin, Editors, John Wiley & Sons, Ltd.: Chichester, England, pages 121-135, 2008.

[20] Gordon, P., A. Kumar, and H.W. Richardson. The Influence of Metropolitan Spatial Structure on Commuting Time. *Journal of Urban Economics*. 26(2): 138-151, 1989.
[21] Hirschberg, D.S. Algorithms for Longest Common Subsequence Problem. *Journal of the Acm*. 24(4): 664-675, 1977.
[22] Zhang, P., et al. Multifunctional nanoassemblies for vincristine sulfate delivery to overcome multidrug resistance by escaping P-glycoprotein mediated efflux. *Biomaterials*. 32(23): 5524-5533, 2011.
[23] Tiedemann, C.E. Classification of Cities into Equal Size Categories. *Annals of the Association of American Geographers*. 58(4): 775-786, 1968.