Sensitivity of DBSCAN in identifying activity zones using online footprints

David W. S. Wong[1*], **Qunying Huang**[2]

[1] George Mason University, USA
[2] University of Wisconsin-Madison, USA

*Corresponding author: dwong2@gmu.edu

Abstract
In mining spatial(-temporal) data for trajectory and activity analyses, a common task is to determine spatial clusters, which may represent activity zones. DBSCAN is a popular clustering algorithm. How the two parameters of DBSCAN that control the clustering algorithm may affect the results spatially has not been thoroughly investigated. This paper reported an incremental effort in conducting a sensitivity analysis by changing the parameter values. Preliminary results show that the two parameters work against each other to a certain degree. Increasing one parameter value (*minpts*) will break up larger clusters into smaller ones, while the other parameter (*eps*) controlled the spatial scale of clusters that can be detected.

Keywords
Spatiotemporal trajectories, spatial clustering, twitter, scale of cluster

I INTRODUCTION

Recent studies in GIScience and spatial analysis often exploited Twitter data, which may be considered as individual-level spatiotemporal data, to understand the mobility patterns of population. Many time-geography studies depict individual trajectories using space-time paths (sets of connected line segments) with data of similar nature to Twitter data. A fundamental issue is that this traditional representation depicts the trajectories of individuals with absolute certainty in space and time regardless if the data were gathered just for one day or several weeks, or derived from data like Twitter.

Twitter data may be regarded as data collected from semi-random spatiotemporal sampling of individual movements. Huang and Wong (2015) proposed a framework to use such data to depict the regular mobility patterns of individuals with certainty levels. The critical process is to identify zones that individuals visited regularly within the 24-hour period. Zones are formed by clustering visited locations within the same temporal window. These zones over different periods are connected with 3D cones to show the spatiotemporal (ST) trajectories. The variable sized ST cones depict the spatial variability of the trajectories and different colour hues on the cones indicate the levels of (un)certainty. A critical step in determining the zones regularly visited by the individuals employed a highly popular clustering method proposed by Ester *et al*. (1996), the density-based spatial clustering of applications with noise (DBSCAN). This clustering method has been adopted in many mining applications of spatial data (e.g., Birant and Kut, 2007; Huang and Wong, 2015; Zhou *et al.*, 2004). How DBSCAN is executed in determining the activity zones will affect the sizes of the ST cones and thus the depiction of the ST trajectories, introducing another uncertainty dimension through the method, in addition to the uncertainty in the data.

DBSCAN requires two parameters: the minimum number of points that can form a cluster (*minpts*) and the maximum distance within which two points belong to a cluster (*eps*). Using different parameter values may obtain different results and thus the spatiotemporal trajectory identification results are likely dependent on these parameter values. Although some studies have suggested appropriate values for *minpts*, these recommendations were data dependent and are not generally applicable. Although conducting a full-scale sensitivity analysis of DBSCAN is warranted to derive some general rules or guidelines in using this clustering method, the objective of this paper is more limited: evaluating the impacts of varying DBSCAN parameters on zone identification and thus ST trajectory variability depiction.

II METHOD AND DATA

Following the recommendations provided by Ester *et al.* (1996), Huang and Wong (2015) set the *minpts* value to 4 and the *eps* value was determined using an iterative procedure. On the other hand, Zhou *et al.* (2004) used 20 meters as the *eps* value as this value approximated the positional error in GPS readings. As these recommended methods in determining the parameter values were based upon specific studies, they are not necessarily applicable to other research contexts, particularly in determining the regular activity zones of individuals based upon locations reported through social media data. Therefore, in this study, we will conduct a limited scope sensitivity analysis using a range of parameter values for DBSCAN to explore how the activity patterns and ST trajectories depiction will vary.

To assess the impacts of varying *minpts*, we use values range from 3 to 10, as Ester *et al.* (1996) suggested using 4. While Zhou *et al.* (2004) suggested an *eps* value of 20 meters, we test its impacts with a much larger range from 10 meters to 60 meters. The maximum *eps* value of 60 meters is probably sufficient to accommodate the locational uncertainty of human activity patterns at the intra-urban scale, but it may not be sufficient for activities conducted in the suburban or even rural areas. Therefore, in choosing our test data, we limited our choices to urban settings. For the sensitivity analysis, Twitter users in Washington, DC and Chicago, IL were used. From the tweets posted between January 1, 2014 and March 31, 2014 with users who used “Washington, DC” and “Chicago” in their profiles, we identified more than 9000 and 17,000 unique users in the two cities. After screening the data with other requirements (e.g., minimum number of geo-tagged tweets has to be 3 or more), 4,442 and 4,088 users from DC and Chicago, respectively, were used. Each of these users posted more than 3 geo-tagged tweets throughout the approximately two-year period (we retrieved the maximum number of tweets, 3,200, for each user).

In addition, we also selected one user likely with his/her residential locations in Washington, DC (we rarely can be definitely sure about the home locations of users) for detailed analysis. Specifically, we want to examine in detail how the clustering results are affected by different *minpts* and *eps* values. Using these data, we test different parameter values using DBSCAN to determine their activity clusters.

III RESULTS OF SENSITIVITY ANALYSIS

The two concerned parameters for DBSCAN control two properties of clusters to be identified. The *minpts* parameter controls how dense point locations will constitute a cluster. The *eps* parameter controls the spatial extent of a cluster. With smaller *minpts* values, more but smaller clusters are expected to be identified. Therefore, using smaller *minpts* values will likely increase the probability of committing type I error (false positive), including random locations that may not be regular activity locations. Using larger *minpts* values will likely produce fewer but larger clusters, likely increasing the probability of type II error (false negative) and collapsing distinctive clusters. On the other hand, how *eps* value may affect clustering results has not been clear, although larger *eps* values may possibly produce smaller numbers of clusters but larger in size.

Figure 1 summarizes the results by varying *minpts* from 3 to 10, and *eps* from 10 to 60 for both Washington, DC and Chicago, IL. When *minpts* increases from 3 to 10, the averaged number of clusters decreases. This result is intuitively expected. When the clustering process requires a larger minimum number of points in order to form a cluster, given that the total number of points to be clustered is fixed, fewer clusters can be formed. This general pattern was found in both Washington, DC and Chicago, IL and is valid regardless of *eps* value. While the impact of changing *minpts* value on the number of clusters seems obvious, how the changes in *minpts* value affect the process is not apparent.

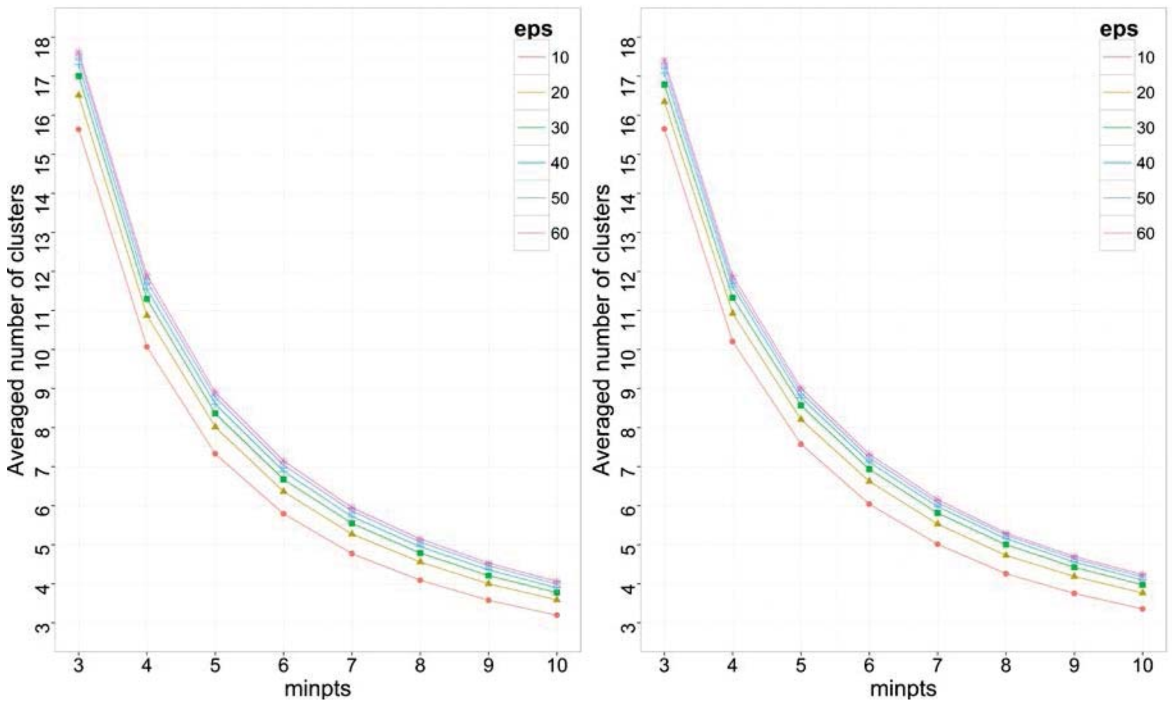
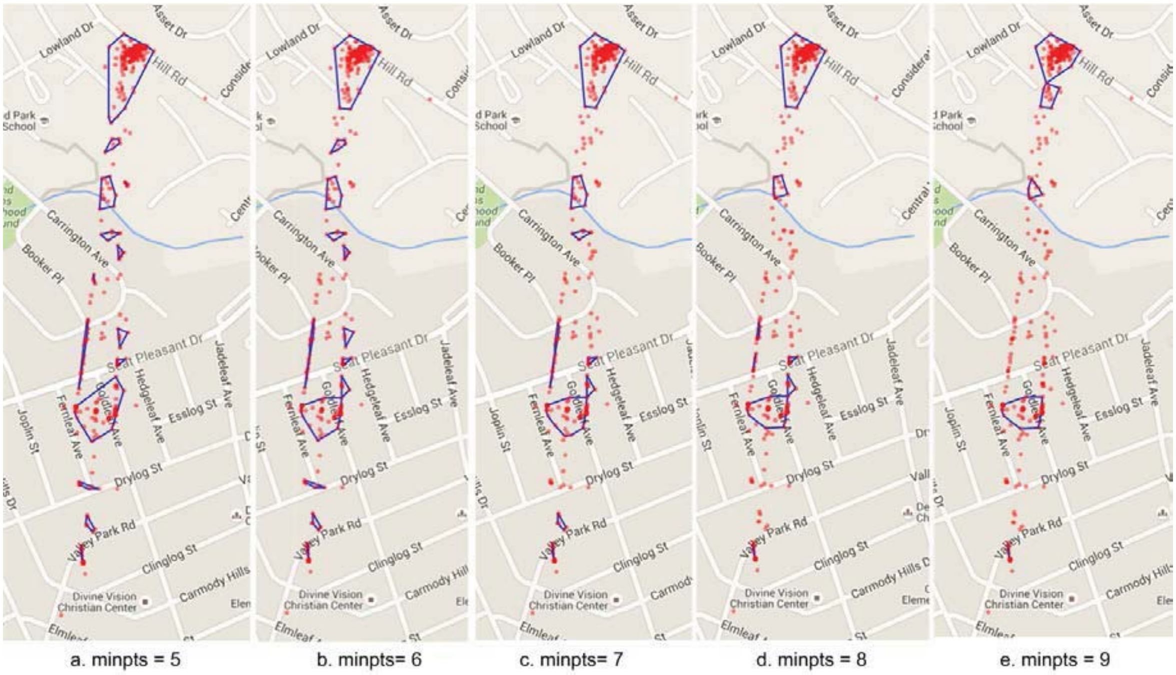


Figure 1: Averaged number of clusters with varying *minpts* and *eps* values (left: Washington, DC; right: Chicago, IL)

Figure 1 also shows that the average number of clusters increases slightly with increasing eps values when *minpts* is kept constant. When *eps* is raised, even points farther away can be included to meet the *minpts* requirement to form cluster. Then, clusters with larger *eps* values may likely be more spatially dispersed, or geographically larger in size. This effect will become more apparently when we examine the clustering results of locations from an individual Twitter user.

We selected a set of locations of tweets from a user to examine the changes in clustering detection when changing the DBSCAN parameter values. Figure 2 shows the clusters identified by DBSCAN using a subset of locations from geo-tagged tweets of a user in Washington, DC. The cluster boundaries were determined using the minimum bounding polygon. The *eps* value was kept at 20 meters, but *minpts* was altered from 5 to 9. In the aggregate analysis described above, we expected that when *minpts* increases, the number of cluster decreases because fewer clusters can be formed as more neighbouring points are required to be reachable within 20 meters for each point to be included in a cluster. Figure 2 shows that when *minpts* value increases, the process ignores smaller cluster and thus fewer clusters are retained. Thus, fewer clusters were formed was not because clusters were competing for points to form larger but fewer clusters. However, the areal extents of clusters shrink when *minpts* increases, quite an unexpected outcome. In fact, larger clusters broke up into smaller clusters when *minpts* value increased (See the cluster(s) on the upper Figures d and e with *minpts* = 7 and *minpts* = 8).



a. minpts = 5 b. minpts= 6 c. minpts= 7 d. minpts = 8 e. minpts = 9

Figure 2: Detected clusters using different *minpts* values and a fixed *eps* value of 20 meters on dispersed locations (The geo-tagged tweets are displayed as red dots and the boundary of a cluster is depicted in blue)

Results reported above show that when value of *minpts* increases, smaller clusters will be removed, but larger clusters will be broken up into more fragmented clusters. This unexpected outcome can be traced back to the design of the DBSCAN algorithm and how *minpts* is used in the clustering process. For each point to be added to a cluster, the point has to reach *minpts* number of points given the *eps* value. So it is possible that a point originally in a large cluster, likely at the edge, with *minpts* value, say 7, could find 7 points, but can no longer find 8 or 9 points within the *eps* value when *minpts* increases to 8 or larger values.

We also examined the varying clustering results when *eps* value changes while keeping the *minpts* value to 4. Figure 3 reports part of the results. With increasing *eps* value from 20 to 70 meters, smaller clusters are merged to form larger but fewer clusters. To a large degree, this general pattern is expected. However, results described in Figure 3 illustrate a fundamental issue in clustering analysis – spatial scale. Cluster detection is a scale-dependent process (e.g., Donnelly, 1978). The *eps* value essentially controls the spatial scale in which the clusters are to be determined. Thus, when *eps* values are small, clusters are small in their spatial extents, and vice versa.

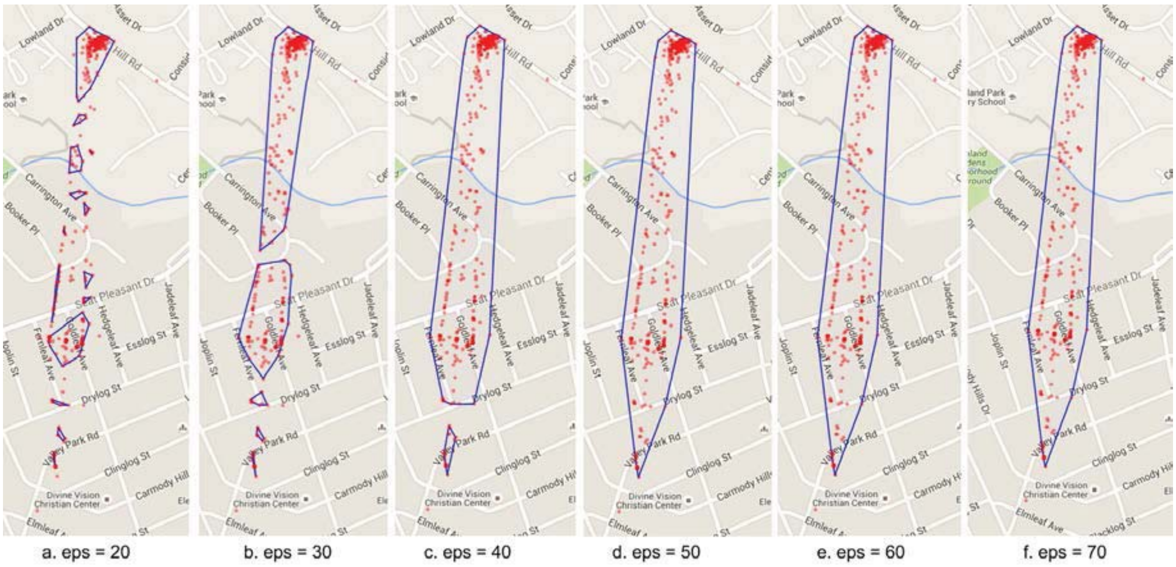

a. eps = 20 b. eps = 30 c. eps = 40 d. eps = 50 e. eps = 60 f. eps = 70

Figure 3: Detected clusters using different *eps* values and a fixed *minpts* value of 4 on dispersed locations

IV CONCLUSIONS

In this short paper, we report only partially results from the sensitivity study. Using larger *minpts* values not only remove smaller clusters, lowering the probability of identifying false positives, but break up larger clusters into smaller clusters. Breaking up larger clusters into smaller ones may not be regarded as "concerning" although the spatial structure of clusters of the entire study area is likely affected. While the results from changing the *eps* value were not surprising, the results highlight the role of the *eps* parameter in determining the spatial scale of cluster detection. Conceptually, determining an "optimal" scale for clustering detection is challenging, not to mention the spatial heterogeneity that may be presented in the data. Taking all these together, we may not be able to qualify the reliability of clustering results using the current tool.

The reported results just focus on the impacts of changing *minpts* or *eps* values, but not simultaneously, or their interaction. On the surface, the two parameters seem to be working against each other: increasing value of *minpts* not only reduces the number of clusters but also areas that fall within the clusters, while increasing *eps* value produces extensive clusters. How the two parameters interact needs to be scrutinized in the future. In addition, the spatial distribution of points to be clustered should also be considered since it may affect the performance of DBSCAN or other clustering detection algorithms.

References

Birant, D., Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering* 60, 208-221.

Donnelly, K. (1978) Simulations to determine the variance and edge-effect of total nearest neighbor distance. In I Holder (ed) *Simulation Methods in Archaeology*, pp. 91-95.

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

Huang, Q., Wong, D. W. (2015). Modeling and visualizing regular human mobility patterns with uncertainty: an example using Twitter data. *Annals of the Association of American Geographers* 105(6), 1179-1197

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L. (2004). Discovering personal gazetteers: an interactive clustering approach. Paper presented at the *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, ACM, pp. 266-273.