Using Density-Based Spatial Cluster Analysis to Study Health-Related Mobility and Community Participation

Eugene Brusilovskiy, Ess Jaraha, Louis A. Klein, Mark S. Salzer

Department of Rehabilitation Sciences, Temple University, Philadelphia, PA, USA

Introduction

Community Participation, as indicated by mobility and engagement in socially meaningful activities, is a central component of health based on the International Classification of Health, Functioning, and Disease (WHO,2001). Global position systems (GPS) technology is emerging as a tool for tracking mobility and participation in health and disability-related research. In an ongoing study, the Temple University Collaborative on Community Inclusion uses AccuTracking Software to collect GPS data from 120 participants with psychiatric disabilities. Participants carried GPS-enabled cell phones which recorded their location every minute for two weeks. This document shows a workflow for processing the data in R. Two density-based spatial clustering algorithms are used: ST-DBSCAN and DBSCAN.

Working with GPS data

There are often inaccuracies and errors associated with GPS data, due to factors including, but not limited to, atmospheric conditions, satellite and reciever errors, and multipath errors. These are reflected in the data as duplicate records, inaccurate coordinates, and missing data. Consequently, the data must be processed before and after clustering.

Analysis

1. Preprocessing

During preprocessing, four main operations are applied:

1. **duplicate records are removed**
2. **outliers are smoothed**
    - an outlier is defined as an observation that is <= 1 minute and > 200 meters from its adjacent observations, where the adjacent observations are <= 2 minutes and < 200 meters from each other. these scenarios correspond to innaccurate coordinates from the GPS. the coordinates of the outlier observation are redefined as the average of its adjacent observations.
3. **lost data is recovered through imputation**
    - coordinates of observations with a time gap <= 20 minutes are linearly imputed.
4. **distance and time fields are calulated**

2. ST-DBSCAN

ST-DBSCAN is a spatiotemporal, density-based clustering algorithm. It clusters observations according to three parameters:

- **eps1**: spatial distance
- **eps2**: non-spatial distance (for the purposes of this analysis, eps2 is time)
- **minpts**: minimum number of points necessary to form a cluster

In this analysis, eps1=200 meters, eps2=20 minutes, and minpts=10 points. Therefore, observations within 200 meters and 20 minutes of at least nine other observations will be clustered. The clusters that result from ST-DBSCAN represent destinations visited by the participant.

3. Postprocessing and DBSCAN

Postprocessing accounts for errors and innacuracies from the STDBSCAN step that were initially caused by GPS errors and innaccuracies.

During postprocessing, four main scenarios are accounted for:

1. **scenario 1**: signal is lost upon arrival at destination, and regained upon departure. arrival/departure points appear as transit points in the data and were not clustered.
    - two adhacent, unclustered observations with distance <= 200 meter and 20 <= time gap <= 720 minutes are assigned to a unique cluster.
2. **scenario 2**: GPS error results in a time gap > 20 minutes, so imputation did not occur. The observation that *follows* the GPS-error-induced time gap was not clustered because of the large time gap. the adjacent observation that precedes the unclustered observation was clustered.
    - the unclustered observation is assigned to the cluster of the adjacent clustered observation if it has a distance <= 200 meters and a 20 <= time gap <= 720.
3. **scenario 3**: GPS error results in a time gap > 20 minutes, so imputation did not occur. The observation that *precedes* the GPS-error-induced time gap was not clustered because of the large time gap. the adjacent observation that follows the unclustered observation was clustered.
    - the unclustered observation is assigned to the cluster of the adjacent clustered observation if it has a distance <= 200 meters and a 20 <= time gap <= 720.
4. **scenario 4**: signal is lost while at a destination. when signal returns, a new cluster is formed at the same destination. this results in two clusters that represent the same instance at a location.
    - two clusters are combined if their centroids are within 200 meters and their time ranges fall within 20 to 720 minutes of each other. the higher cluster id will be preserved.

After postprocessing, stdbsan-clusters are clustered spatially using the dbscan package. The clusters resulting from DBSCAN clustering represent repeated destinations.

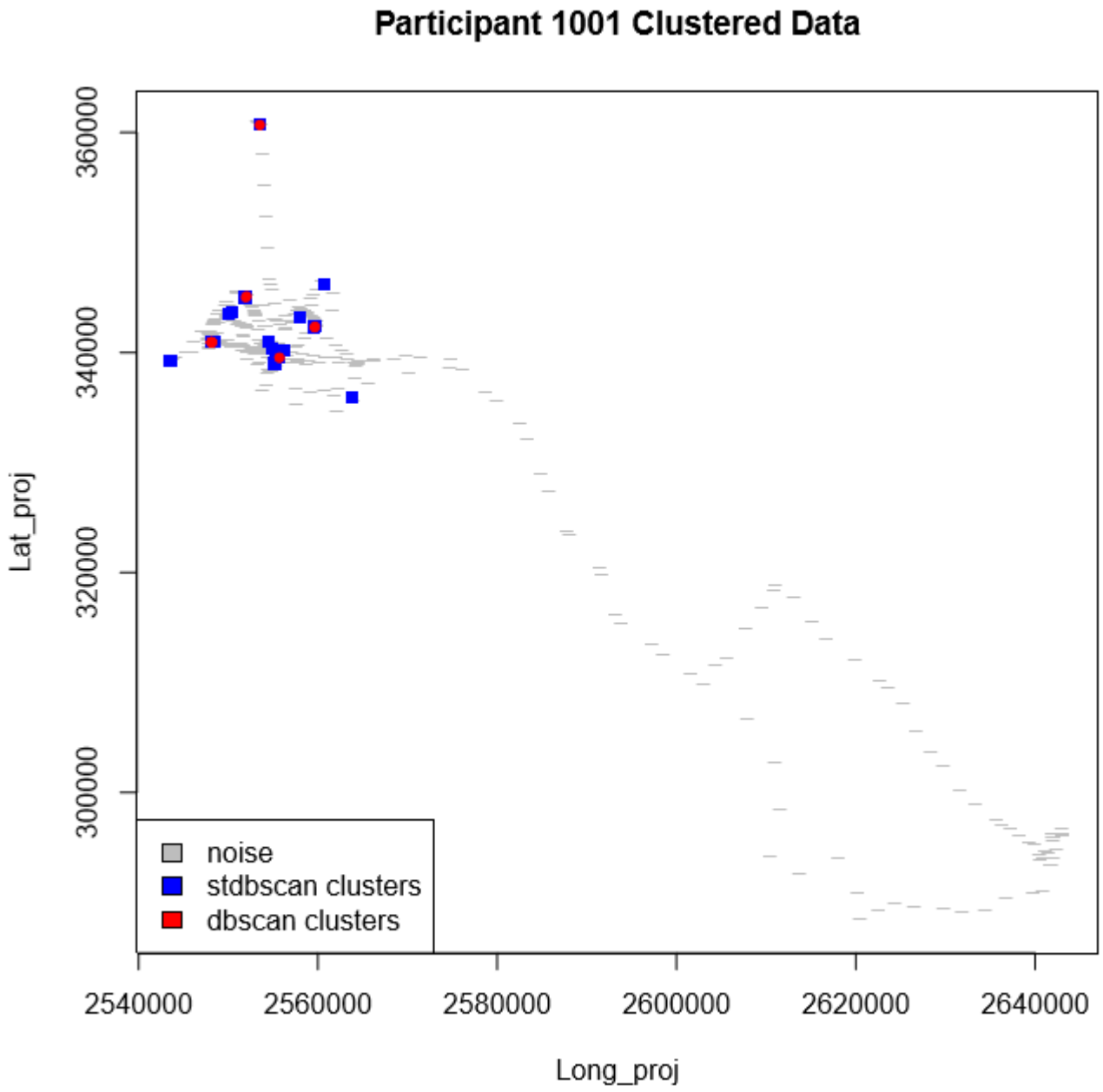
4. Data Visualization

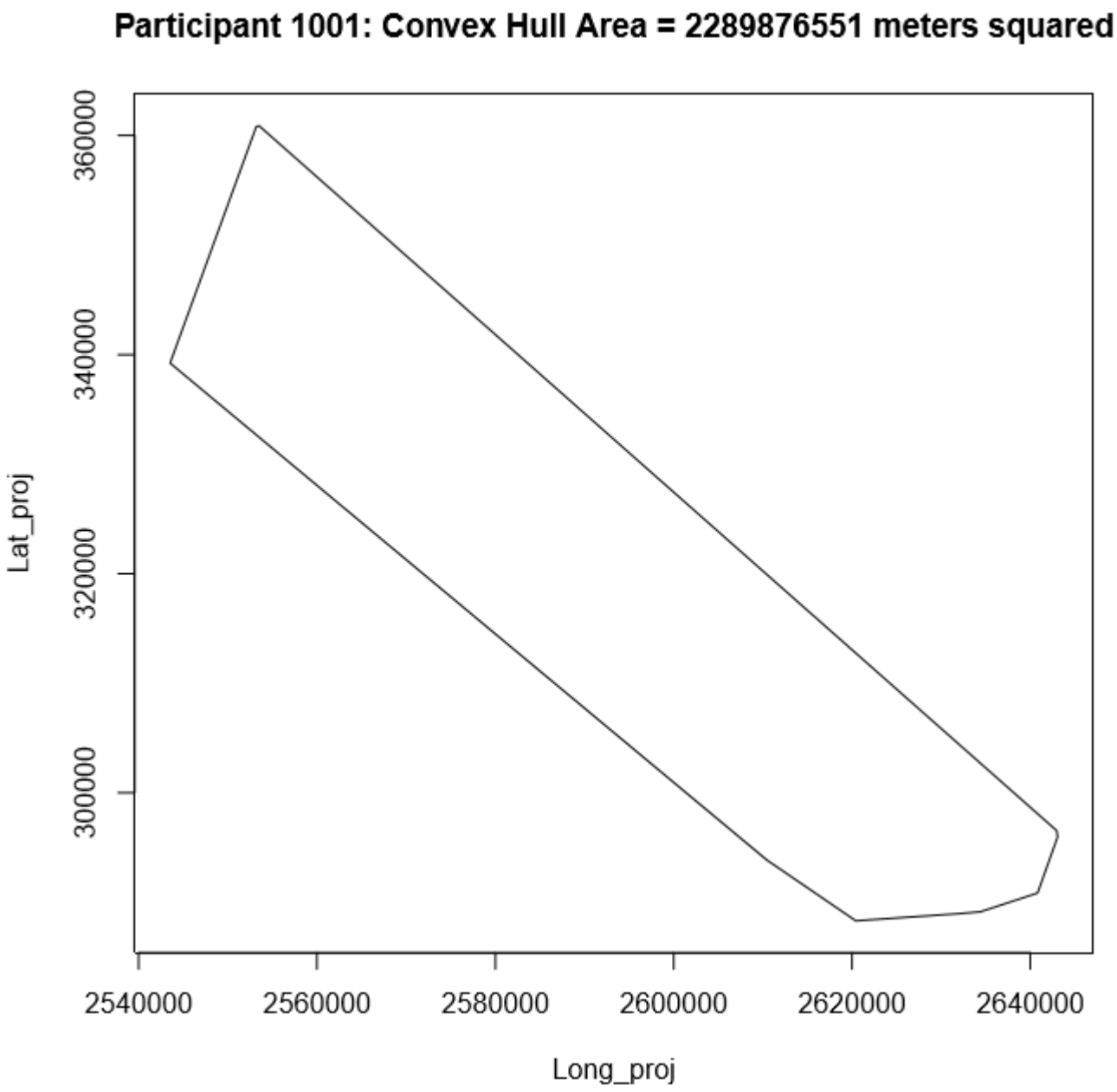
Three functions were developed to visualize participant data.

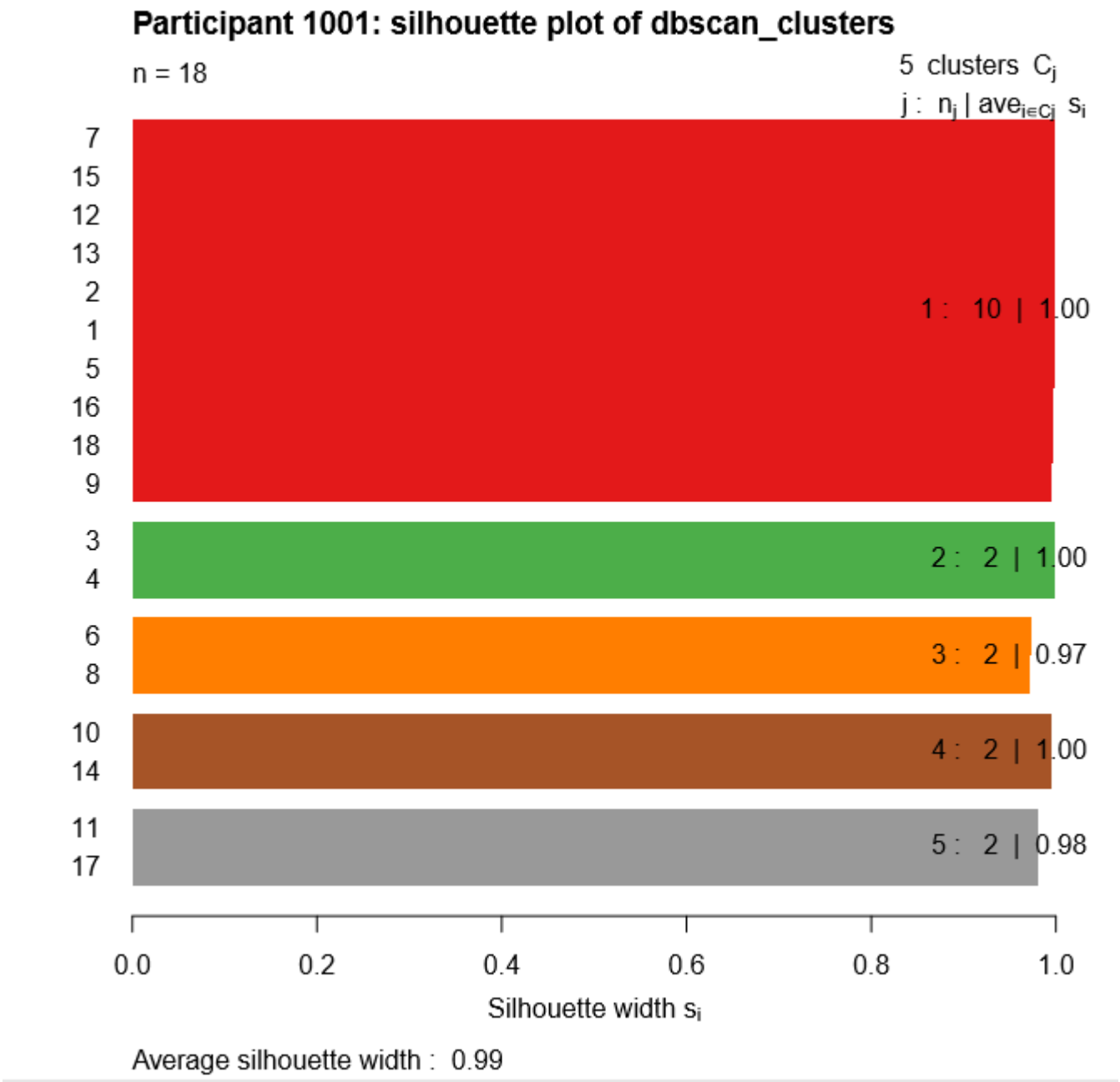
1. **cluster_plot**: scatter plot of **stdbscan clusters**, **dbscan clusters**, and non-clustered points for a participant
2. **chull_area_plot**: polygon plot showing the **convex hull** and area of the convex hull for a participant
3. **silhouette_dbscan**: bar plot showing results from a **silhouette analysis** on dbscan clusters
    - a silhouette analysis is an internal validation method which assigns a score (s_i) to clusters based on the distance between points within a cluster and the distance between clusters. more information can be found in the silhouette function documentation (https://www.rdocumentation.org/packages/cluster/versions/2.0.9/topics/silhouette).

All functions generate pdfs and require two types of input:

1. path(s) to the files containing the necessary data
2. path to the pdf to be created




Participant 1001: Convex Hull Area = 2289876551 meters squared
Lat_proj
300000
320000
340000
360000
2540000
2560000
2580000
2600000
2620000
2640000
Long_proj



References

World Health Organization, 2001. ICF: International Classification of Functioning, Disability, and Health (Geneva).