

On the Quest for Representative Behavioral Datasets: Mobility and Content Demand

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore

► **To cite this version:**

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore. On the Quest for Representative Behavioral Datasets: Mobility and Content Demand. CHIST-ERA Projects Seminar 2016, May 2016, Bern, Switzerland. hal-01323917

HAL Id: hal-01323917

<https://hal.inria.fr/hal-01323917>

Submitted on 31 May 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1. OBJECTIVES

Understand correlations between human mobility and data demand.

Objectives:

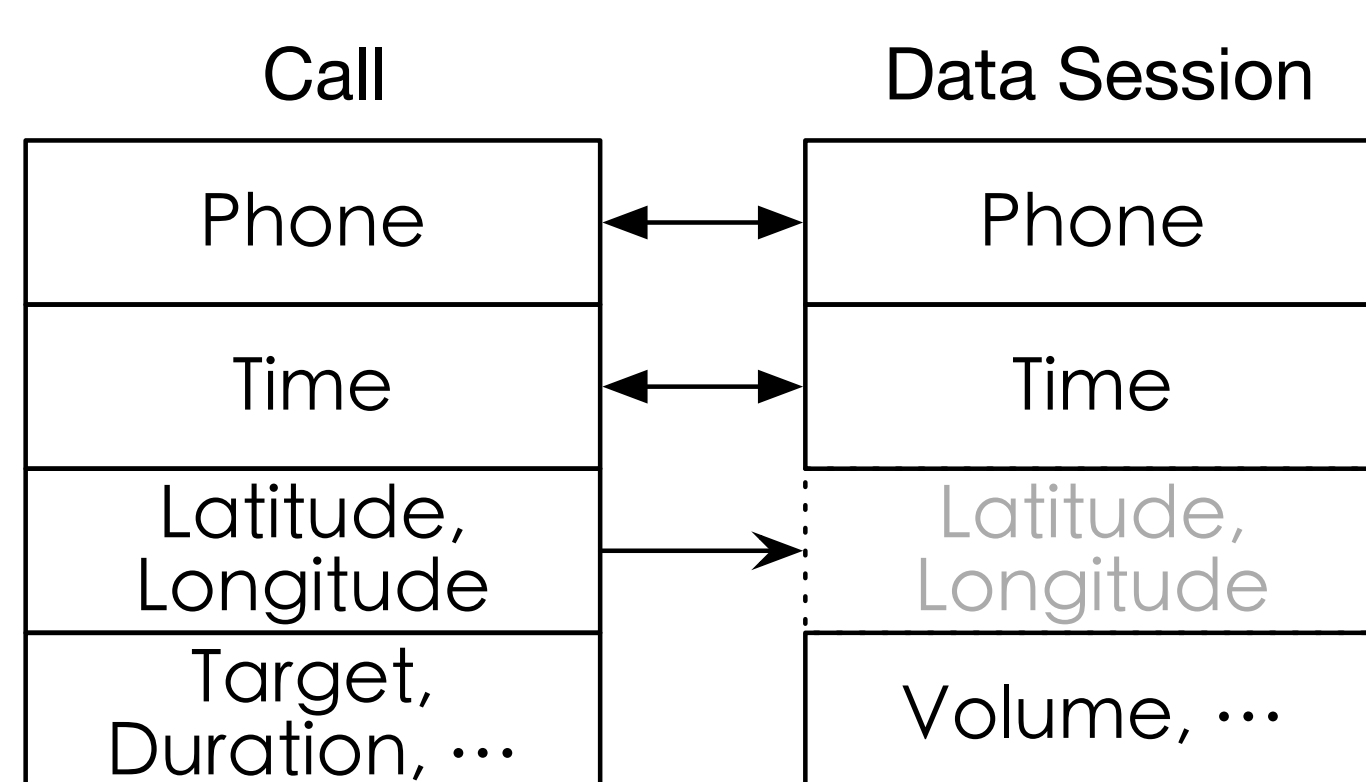
- Extract fully-featured dataset from cellular traces;
- Construct mobility: identify Home locations, and estimate trajectories;
- Characterize user behaviors in terms of time, space and volume (data).

2. DATASET DESCRIPTION

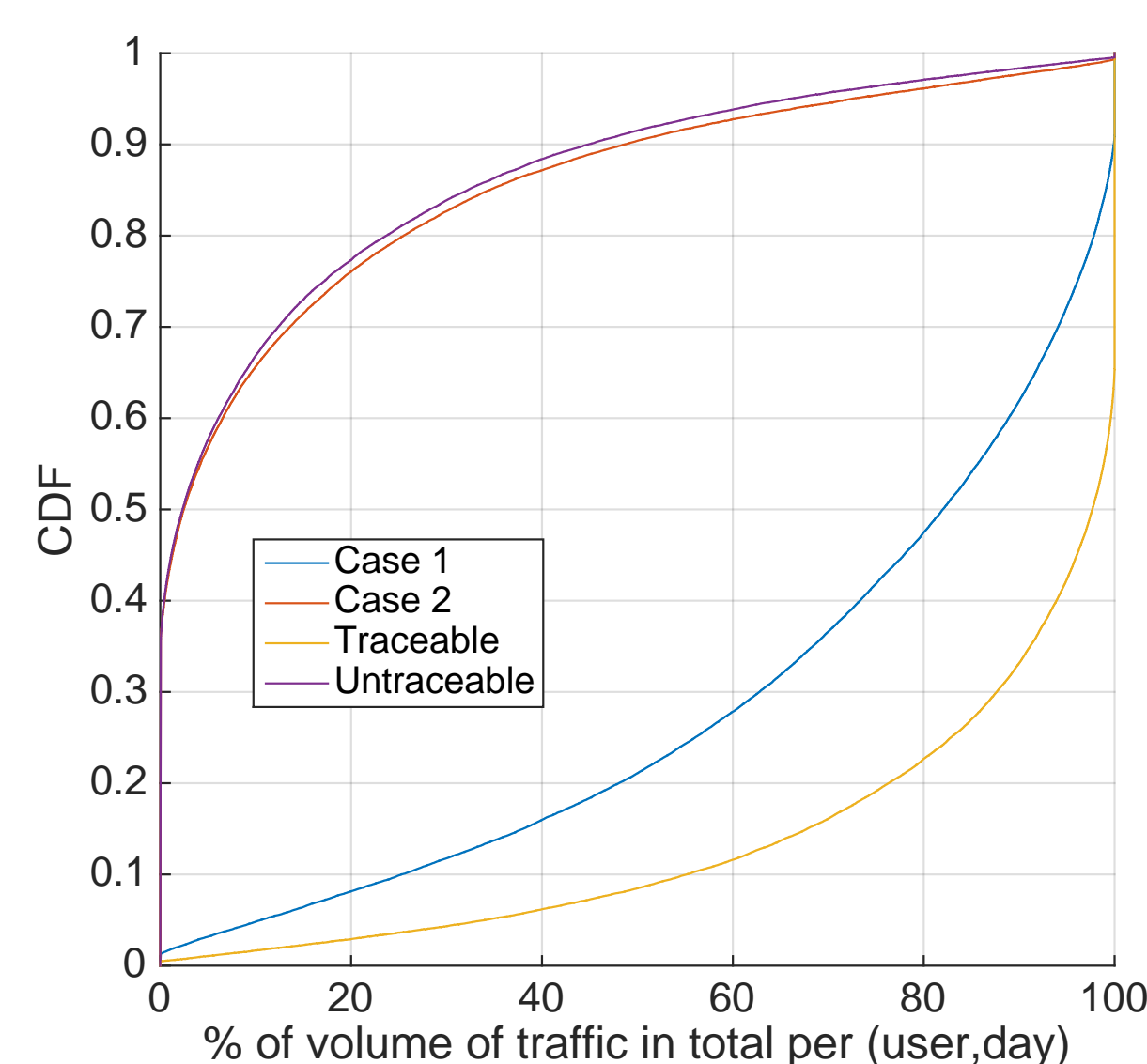
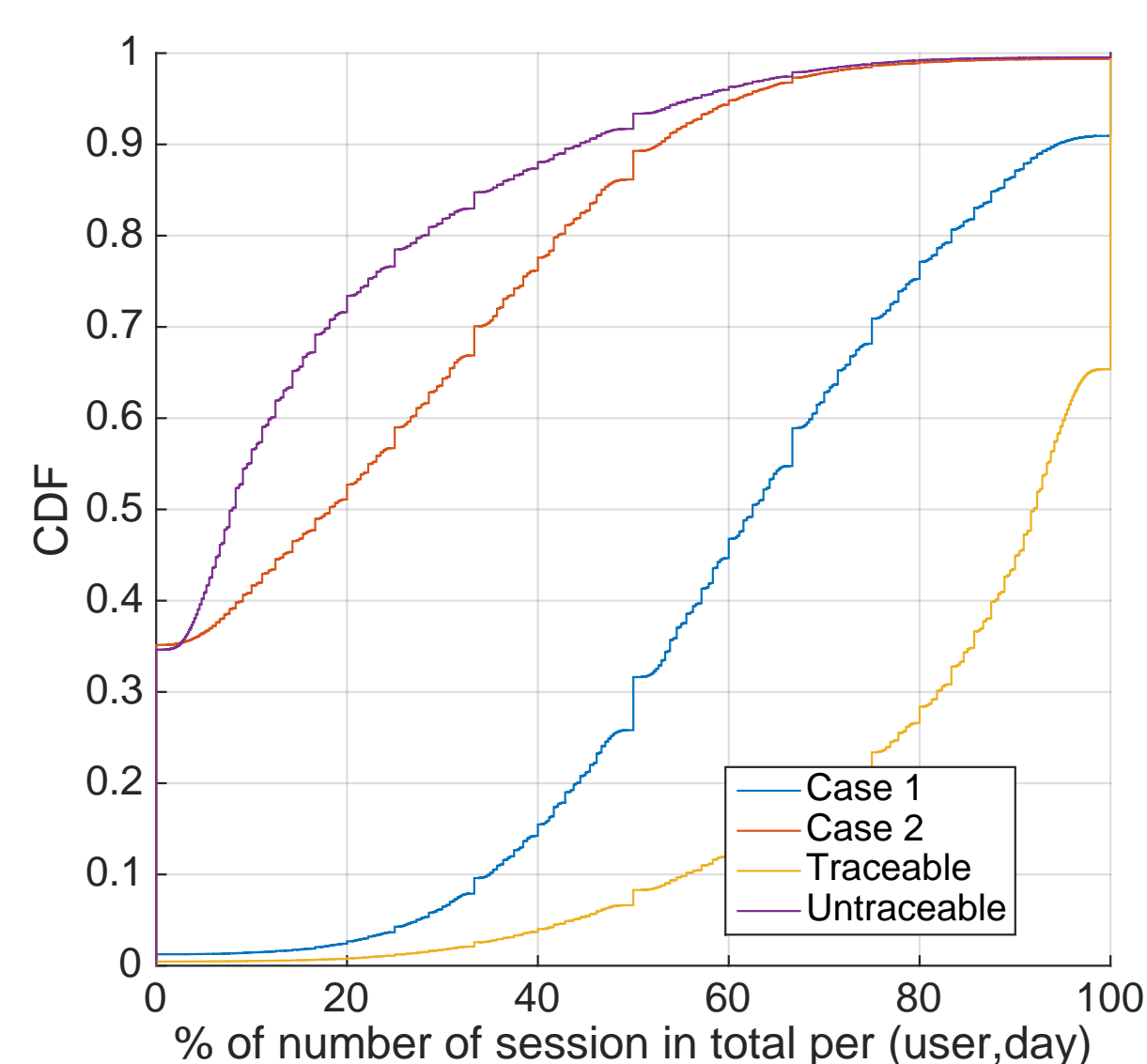
17,366 subscribers are extracted by applying a series of filters on cellular traces, consisting of 2,398,392 calls and 954,737 sessions in 4 weeks.

Session Location Estimation

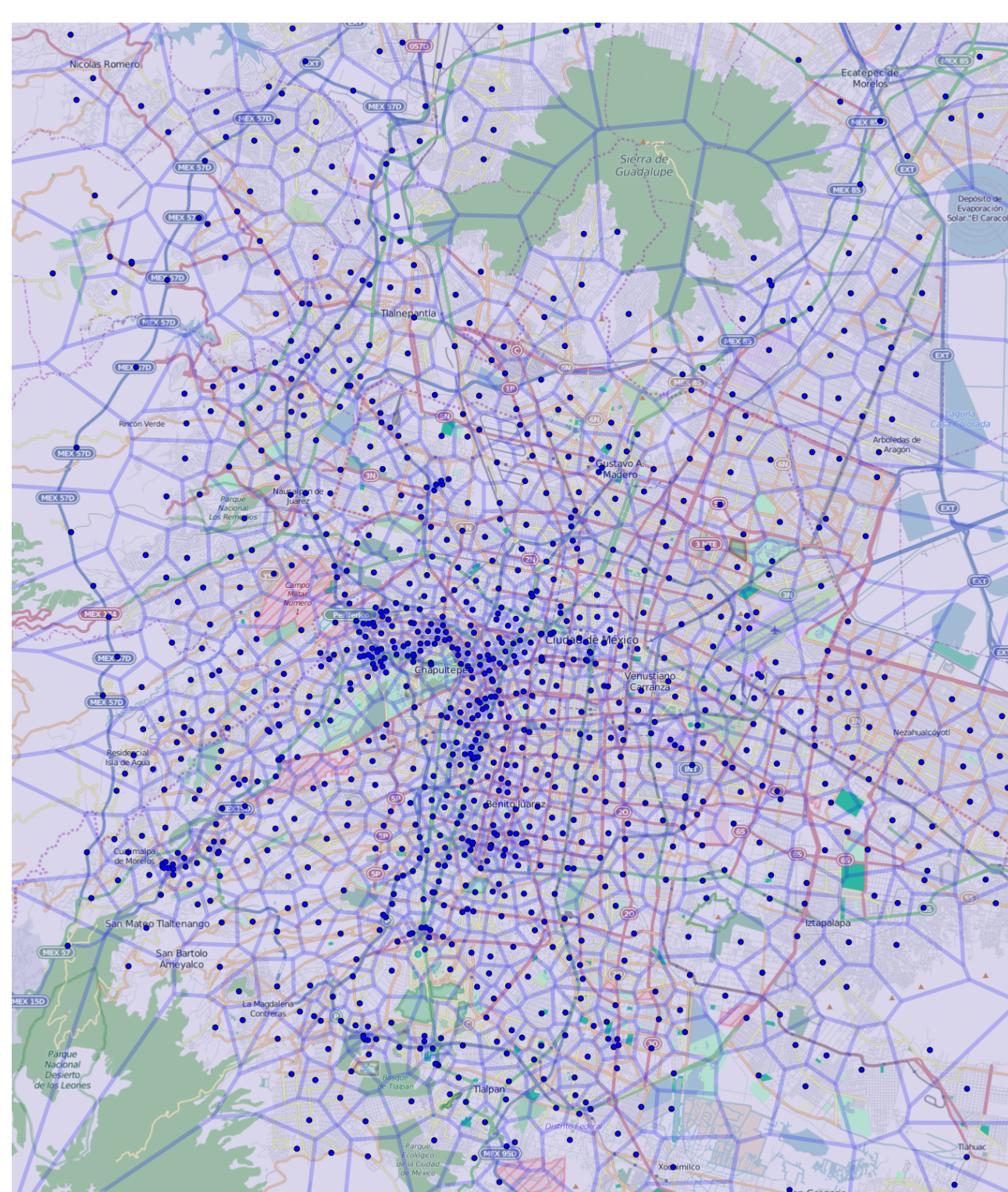
Using call detail records, we infer users' locations for data sessions.



Based on trajectory estimation, at least 70% of sessions (80% of volume) are traceable for 80% of subscribers.



Decomposition of Cells in Mexico City



Cellular Cells (Voronoi)

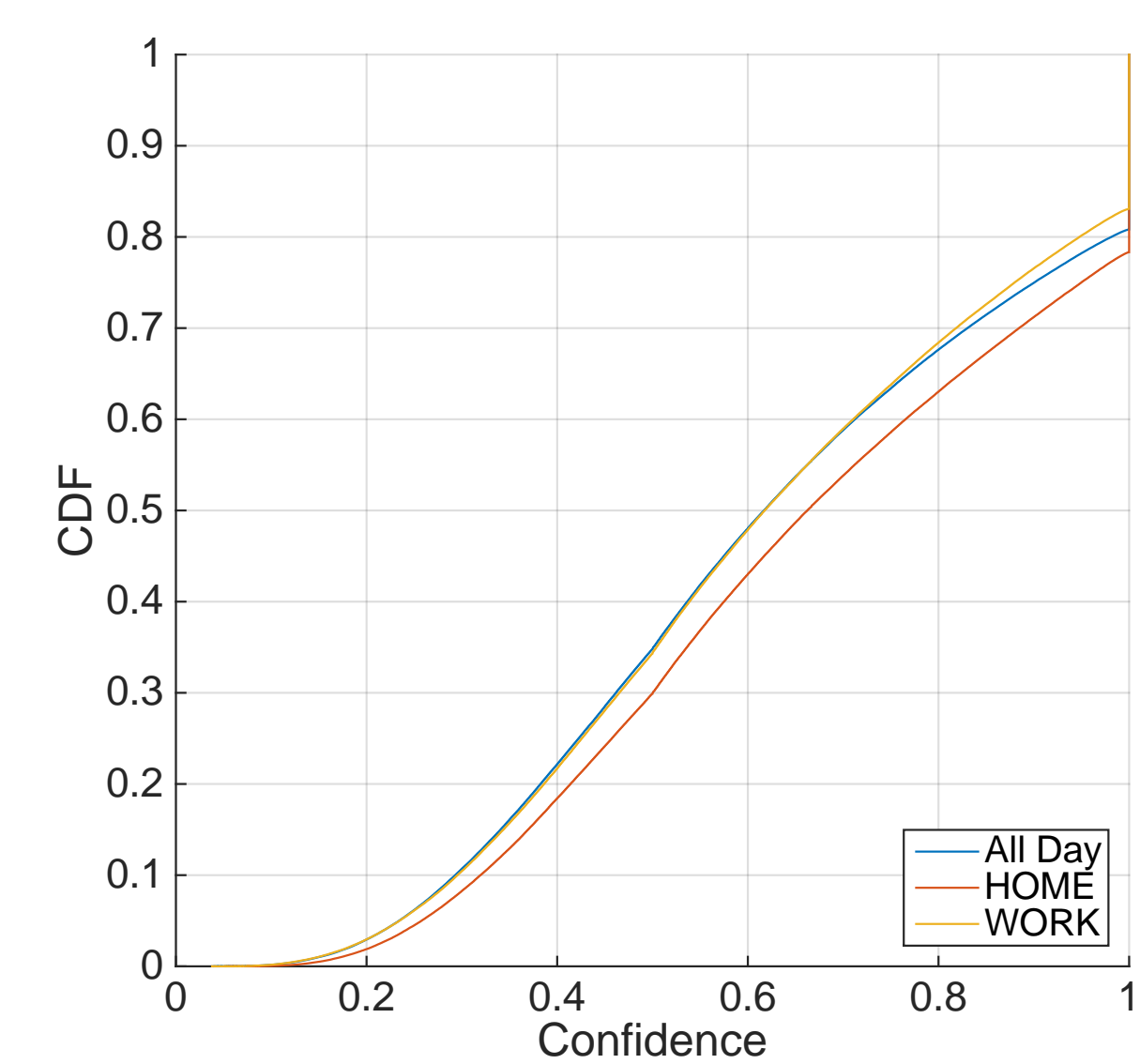


Population Density

3. ANALYSIS

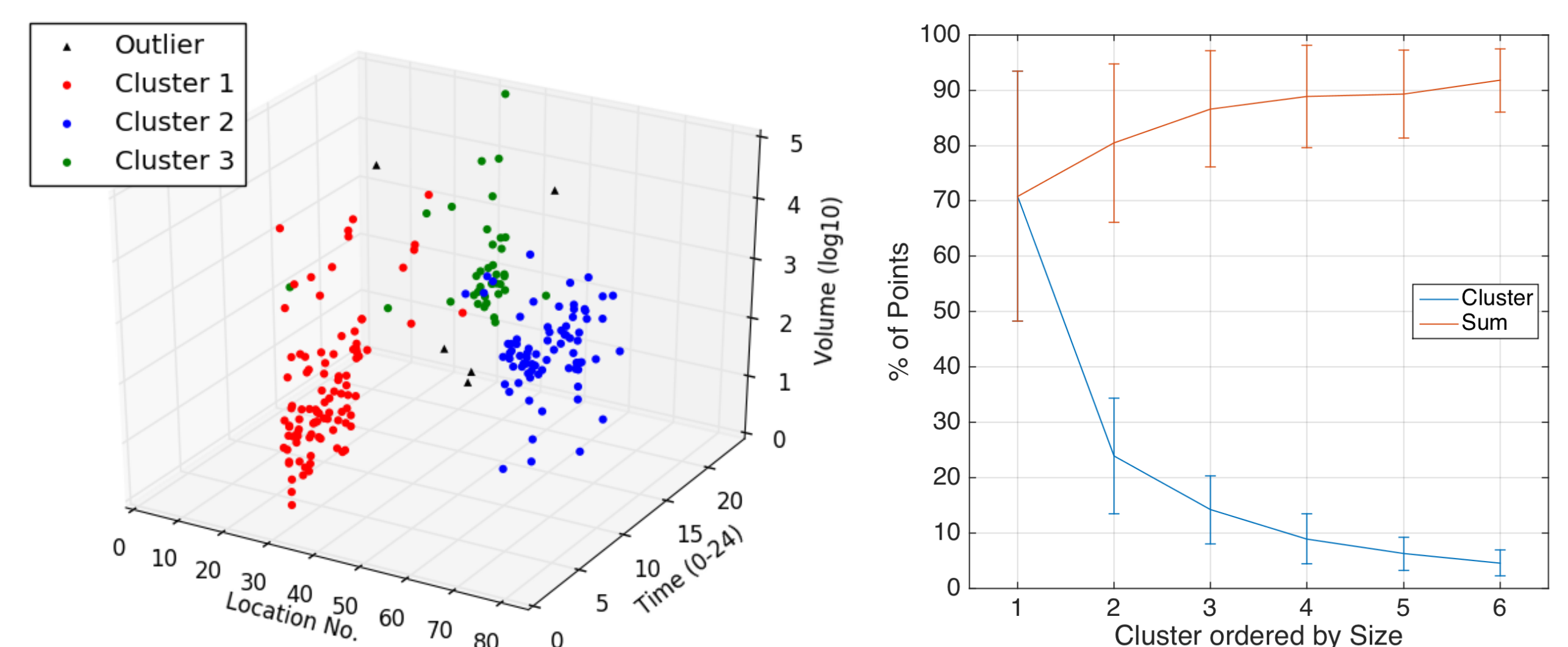
Home Location Identification

For each user, the most visited cell between 10pm and 7am is identified as his/her *HOME*.



Clustering Sessions

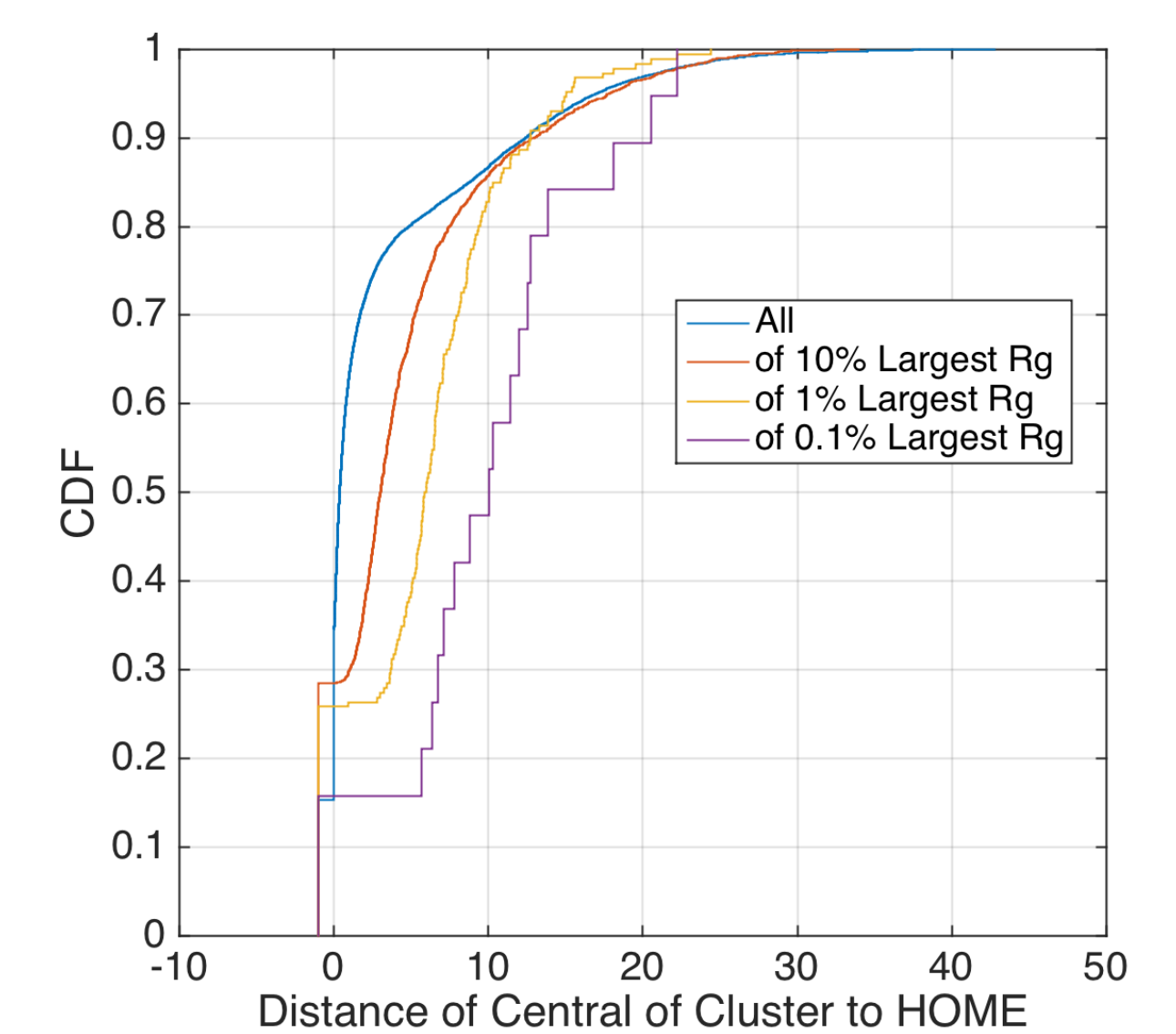
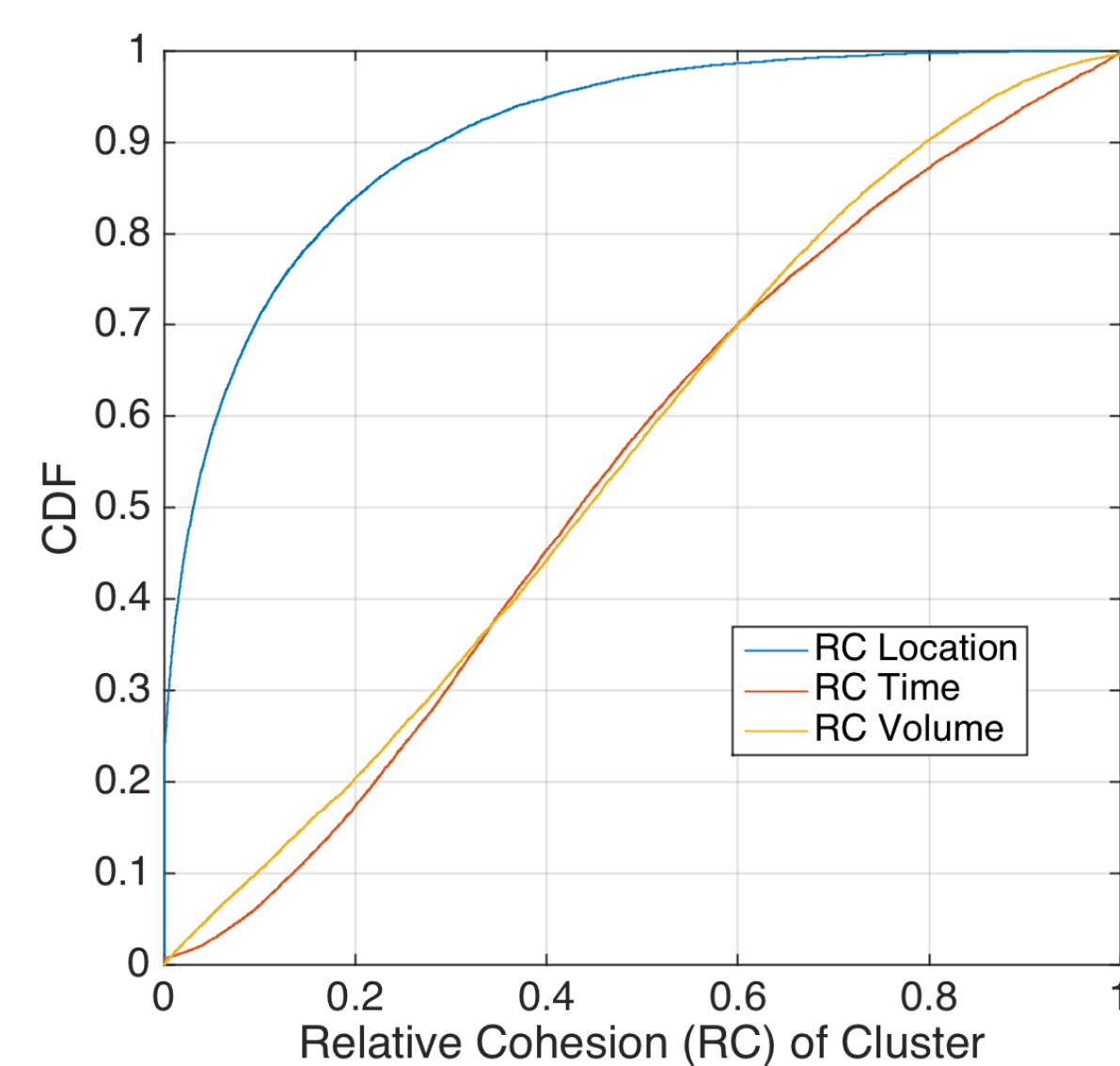
Sessions (space, time, volume) are clustered by DBScan.



For each cluster, **Relative Cohesion** is calculated as a function of time, space or volume, respectively.

$$RC^{(*)} = \frac{\sum_{p \in C} dist^{(*)}(p, c)^2}{\sum_{p \in C} dist(p, c)^2} \quad (1)$$

$$RC^{(loc)} + RC^{(time)} + RC^{(vol)} = 1 \quad (2)$$



4. FUTURE WORKS

- Estimate trajectories' incomplete information;
- Model volume demand;
- Characterize content demand;
- Link and predict mobility (important locations) and volume/content demand simultaneously.