Estimating Dwells using Mobile Phone Network Data

Author: Yubo Wang | **UCL Geography** – MSc Social and Geographic Data Science
UCL Supervisor: Dr Stephen Law | **Partner Institution Supervisors:** Cristóbal Montt

An ESRC Data Investment

Consumer Data Research Centre

Background and Motivation

Understanding human mobility requires a focus on "dwells," which are critical for identifying where people spend significant amounts of time. Call Detail Records (CDR) are valuable for studying dwells due to their extensive population coverage and easy accessibility.

However, several challenges exist:

1. **Voronoi Method:** Inaccurate location estimates and strong assumption.
2. **Machine Learning Algorithm:** Inefficient for large CDR data; requires ground truth calibration.
3. **Geometric Functions:** Data misinterpretation in modifiable areal unit problem (MAUP) and high computational complexity.

Methodology

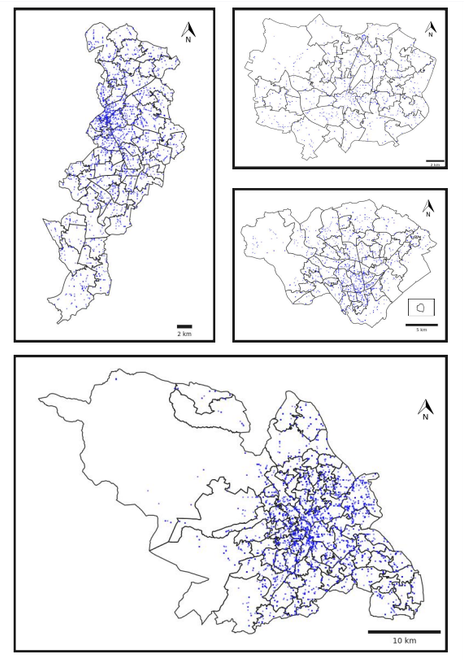
This study focuses on algorithm development through following key stages:

1. **Estimating cell range**: Using flicker events to define cell range and coverage areas.
2. **Filtering data and location identification**: Identifying new locations after time window process.
3. **Generating "Dwell" tables**: Recording dwell times and location, assigning the spatial index.
4. **Inferring home and workplace locations**: Applying the H3 grid system for location detection and multi-scale results comparison.

Each step has been validated to ensure the approach's reliability and effectiveness.

Research Area and Data

- **Manchester/Cardiff/ Sheffield/Coventry** with 10 km buffer
- **June 1, 2024** to **June 15, 2024** (2 weeks)
- **14 billion records** from **6.5 million users** across **27 thousand cells**
- **About one-third of UK Users**



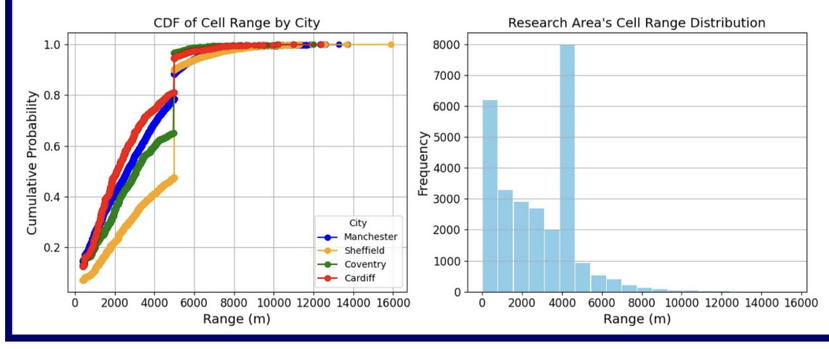
Main Process and Result Analysis

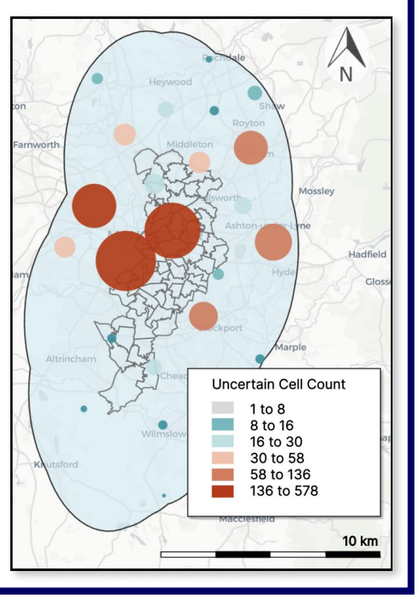
Identify the Flickers and Cell Range

We analyzed "flickers" in the CDR data, which are frequent changes between two cells, to determine the range of each cell. Identified the point where the frequency of flickers changes most sharply to define each cell's effective range.
Using the azimuth and beam width, we created coverage areas for each cell.

Minimum range: **400 meters**
If data is insufficient: **5000 meters**

80% of the cells have an effective range, with uncertain cells more densely located in urban areas. The charts display the distribution of cell ranges and the areas with uncertain ranges.

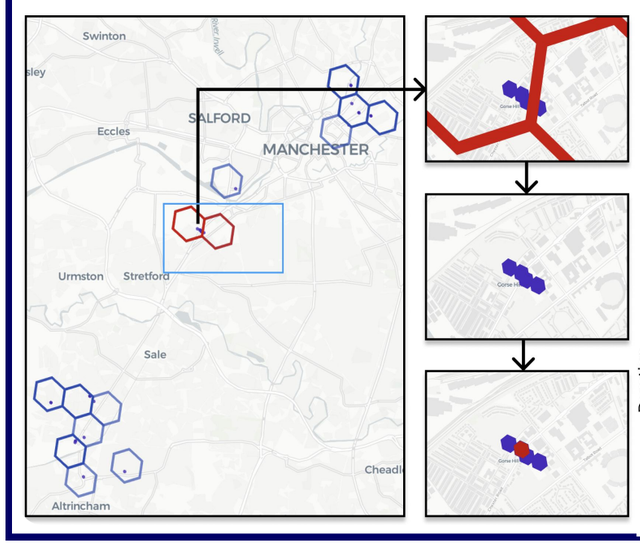




Identify the Home and Workplace

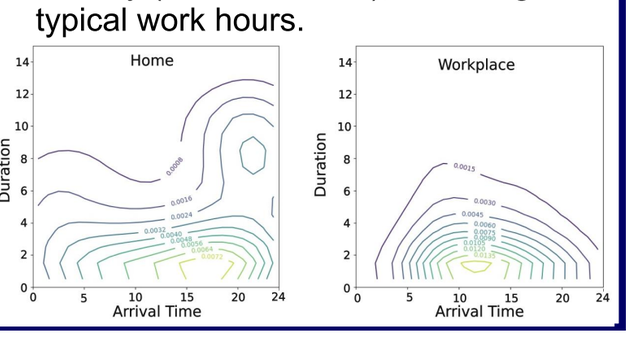
Dwell points are categorized by weekday/weekend and day/night, and assigned high- and low-resolution **H3 Hexagonal grids**.

Home and workplace are detected using **contextual and temporal patterns**: the home is identified from nighttime data, and the workplace from daytime and weekday data. Adjustments are made for night-shift workers as needed.



Home dwell are longer in the evening and early morning (18:00 to 09:00) and shorter during the day.

Workplace dwell primarily occur during the day (09:00 to 17:00), reflecting typical work hours.

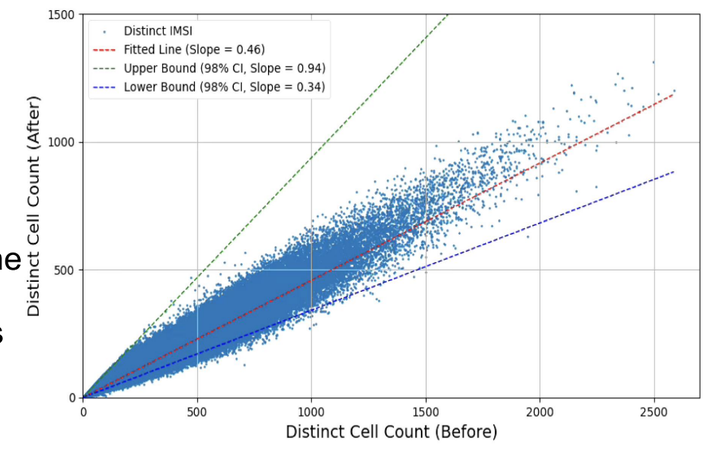


Raw Data Merging and Filtering

The CDR data was filtered using 1-, 3-, and 10-minute time windows to reduce high-frequency flickers. The 1-minute and 10-minute windows reduced distinct cells by 72% and 18%, respectively, but introduced excessive noise or risked losing important information.

In contrast, the 3-min window achieved a **54% reduction** in distinct cells, preserving data quality.

The 3-min time window reduced the initial **13.95 billion to 2.58 billion** records. Decreased I/O operations and enhanced computational efficiency.

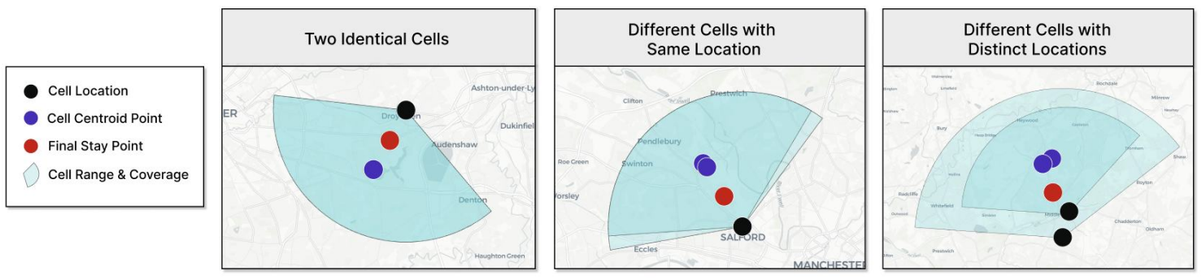


Identifying Final Dwell Locations

After cleaning, the data are deemed reliable. Consecutive data are checked by **Time**, **Cell ID**, **Distance**, and **Speed** to identify significant location changes. Intermediate movements are ignored, preserving only new positions and dwell times.

The final location, representing the dwell, is calculated as the **geographic average** of four points: the cell locations at the start and end of the records and their respective centroid locations.

In all **three scenarios** of start and end records, the resulting final location lies within the coverage areas of both cells and is close to the cell positions.



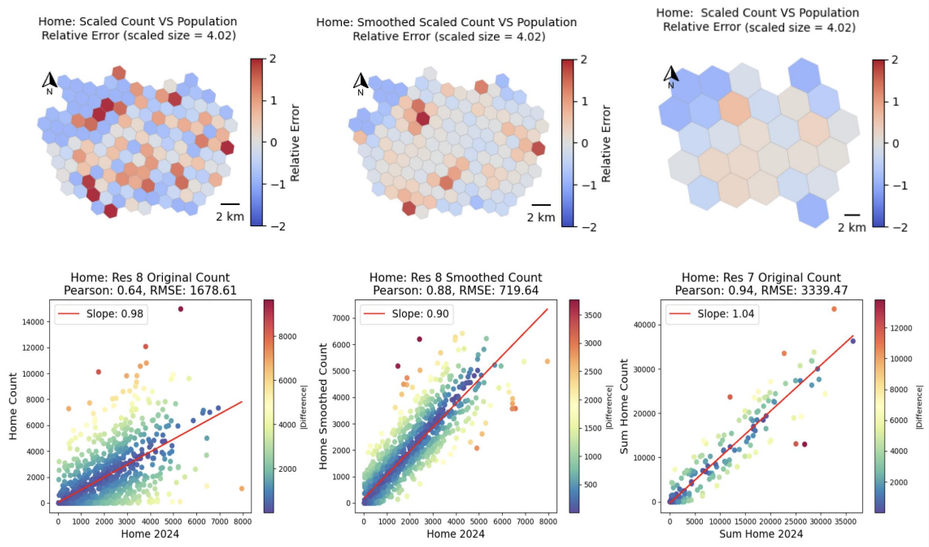
Accuracy Assessment Based on Census Data

Aggregation using **MSOA** boundaries showed low correlation due to uncertain cells and irregular edges, with Pearson coefficients of **0.29** and **0.66** for home and workplace. Therefore, we used H3 Hexagonal grids for better accuracy.

To address MAUP and compare projection methods, we tested three aggregations: Res8, Smoothed Res8, and Res7.

Result show:

- **Res8**'s home projection had a **Pearson 0.64, higher than the MSOA** method.
- **Smoothed Res8** had the **lowest RMSE** of 720.
- **Res7** achieved the **highest Pearson** value of 0.94.



Value of the Research

1. Demonstrate the effectiveness of methods **beyond Voronoi diagrams** for calculating cell range, achieving 80% valid data from two weeks.
2. Validate the use of a 3-minute time window and flicker filtering to **reduce data volume** in detecting dwells using CDR data.
3. Confirm the geometric validity of using the geographic average as the **rational final user location.**
4. Introduce a spatial index to infer home and workplace locations, **beyond the clustering algorithms**. Present results at different resolutions to meet various analytical needs and eliminate the MAUP issue.

Future experiments should use **larger datasets** (rural/ urban) over extended periods (3 months or more) and require **individual-level ground truth** for validation.