GPS Data Analysis for Individual Dwells

CDRC Programming Assignment

Prepared by: **Setareh Nouri**

1. Introduction

The objective of this analysis is to extract insights from Global Positioning System (GPS) data to estimate individual dwells and explore patterns, durations, frequencies, and any limitations related to the data or dwell periods. Dwell locations are defined as sequential observations with little or no movement over a certain period. Table 1 presents a sample of observations from the dataset.

2. Exploratory Data Analysis

The dataset includes individual IDs and 2-dimensional geographic coordinates (latitude and longitude) along with timestamps, which are processed to identify dwell periods. **Table 1** presents a sample of observations from the dataset.

index	user_id	datetime	lat	lon
0	00F70625-4B30-4B4F-A0E3-A5CD9474E34F	2018-01-09 08:49:29	51.5040166231844	-0.0864545342253767
1	00F70625-4B30-4B4F-A0E3-A5CD9474E34F	2018-01-09 08:52:01	51.5051924173689	-0.0907412849307962
2	00F70625-4B30-4B4F-A0E3-A5CD9474E34F	2018-01-09 08:52:32	51.5040983632687	-0.0953212659806901

Table 1: Sample of observations from the dataset

Analyzing the recorded timestamps shows that the GPS data were collected between January 8, 2018, and January 19, 2018, spanning 10 workdays and capturing observations throughout the day. The dataset contains 31,606 unique IDs, with a wide variation in the number of observations per user, ranging from 20 to 3,430 records. **Figure 1** presents a histogram illustrating the distribution of observations recorded for each user.

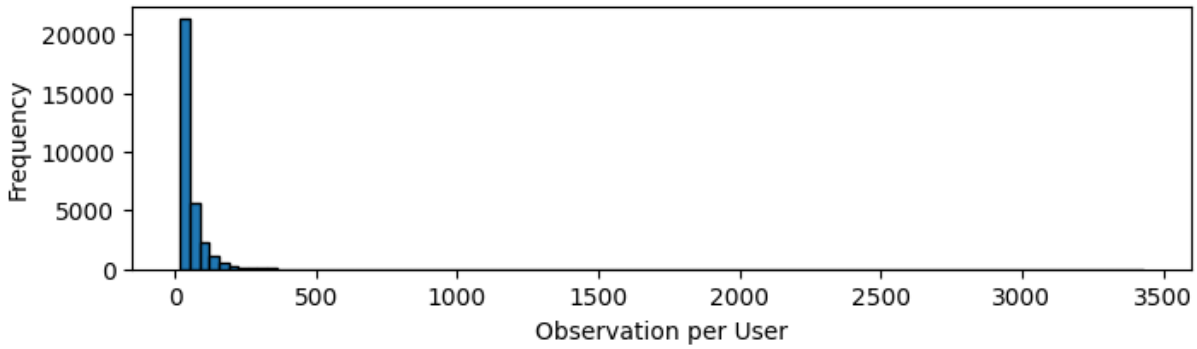

Figure 1: Distribution of observation count per user

The GPS data is located within the geographical coordinates ranging from approximately 51.4825 to 51.5066 latitude and -0.1213 to -0.0764 longitude, corresponding to central London. These statistics are provided in **Table 2**, presenting the minimum, maximum, mean, and standard deviation values of recorded coordinates.

index	latitude	longitude
min	51.4824891	-0.1213403
max	51.5066432	-0.0764079
mean	51.4989555	-0.0983619
std	0.006003	0.011862

Table 2: Statistical description of coordinates

In order to analyze the average movement of users, pairwise distances for individual IDs were computed. It reveals that the mean distance between consecutive GPS points is approximately 0.0045 kilometers. Histogram in **Figure 2** displays the distribution of average distances between consecutive GPS points for users. One observation is that a few individuals (5 users) have remained idle for the entire duration of the data collection period.

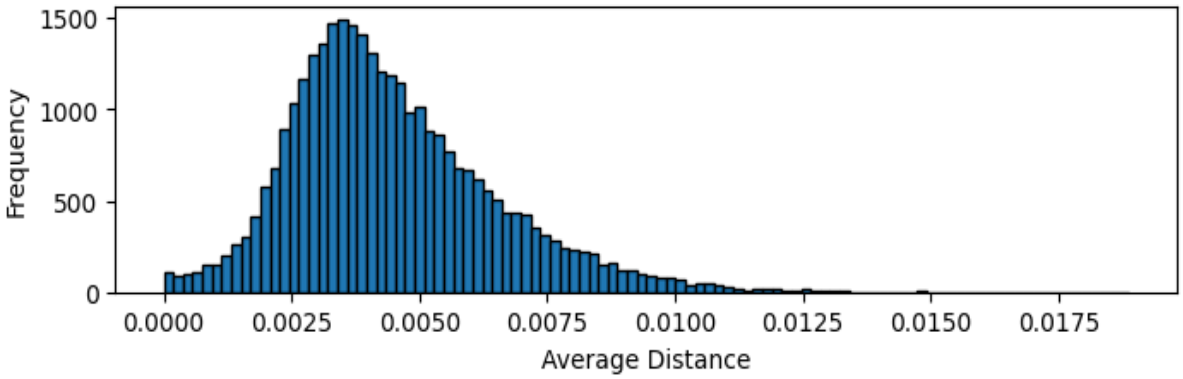

Figure 2: Distribution of users' average movement

3. Methodology

3.1. Criteria

Defining criteria for identifying dwell periods involves determining time and distance threshold limits. Initially, the threshold of a 50-meter radius and a minimum duration of 15 minutes were selected intuitively based on domain knowledge of the data. To refine the threshold, I considered the mean distance between consecutive points for each user and also employed an iterative approach.

3.2. Algorithm

To identify the individual dwells, the clustering approach is used. Dwells are defined as clusters within the specified thresholds. In this work, the DBSCAN algorithm due to its robustness in handling noise and varying density of data points. Also, unlike traditional clustering algorithms, this method does not require presetting the number of clusters, making it adaptable to irregularly shaped datasets. The Algorithm is implemented to cluster GPS data points based on their spatial proximity, with a maximum distance threshold of 0.0045 kilometers (the mean distance between consecutive GPS points). This parameter is used to adjust the epsilon parameter for DBSCAN. The minimum number of samples of is set to 4 (suggested by Sander[1] for 2-dimensional data).

[1] Sander, Jörg, et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." Data mining and knowledge discovery 2 (1998): 169-194.

4. Results

The analysis resulted in the identifying 12,475 individual dwell instances across all users. The map visualization shows the geographical distribution of individual dwellings differentiated by their id (see **Figure 3**), providing insights into the spatial patterns of dwell locations.

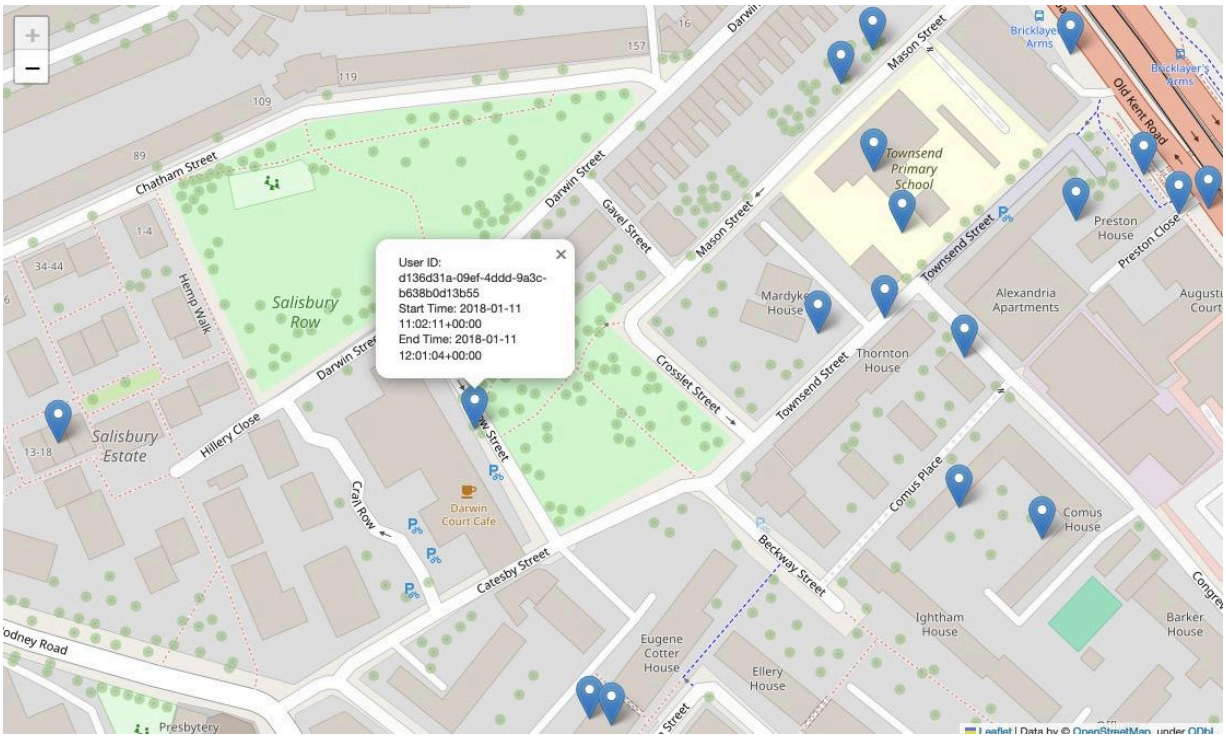

Figure 3: Sample visualization of an individual dwell

The identification of individual dwell instances enables a deeper understanding of user mobility patterns and behavior. The spatial distribution of dwellings highlights areas of high activity or concentration, which may have implications for urban planning, transportation, or location-based services. For example, as depicted in **Figure 4**, the dwell analysis reveals numerous instances of dwell events along Walworth Road in London, known for its diverse array of stores. Also, analysis of dwell durations and frequencies can aid in identifying frequently visited locations or points of interest for users, which is not the purpose of this assignment.

Figure 4: Visualization of individual dwells on Walworth Road in London

According to the histogram in **Figure 5,** the analysis of dwell instances reveals a varying frequency distribution of dwell instances across users, indicating differences in mobility patterns and activity levels. The duration of dwell instances ranges from 15 minutes to several hours, with an average of 144 minutes and mode of 216 minutes. These values and number of identified dwells could vary based on the selected parameter of the algorithm.

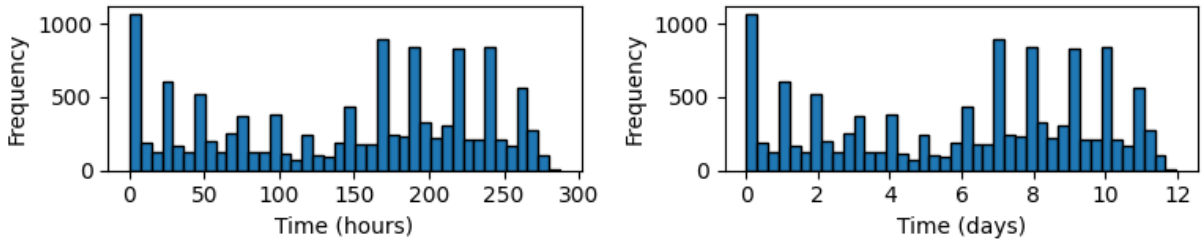

Figure 5: Frequency distribution of individual dwells

5. Challenges and Recommendations

One of the challenges in this assignment is defining the criteria for identifying dwell thresholds, i.e., duration time and distance. Methods like grid search or Bayesian optimization can be applied to find the optimal values for these parameters based on the final application of the results. The end use of the analysis result plays a crucial role in determining suitable criteria for threshold optimization in dwell analysis. Different applications may prioritize different aspects of the analysis result, and the chosen criteria should align with the specific objectives and requirements of the application. For example, in a recommendation system, accurate identification of dwell events helps understand user preferences. In urban planning, it aids in optimizing resource allocation by pinpointing areas of high pedestrian activity. We can further investigate personalized thresholds for users based on their behavior by considering diverse criteria for dwell detection. For instance, individuals traveling at faster speeds might exhibit shorter dwell times, whereas slower speeds could necessitate longer durations to be classified as a dwell.

Another challenge is the high computational complexity of this analysis due to the large scale of the temporal spatial data and the need for processing at the individual user level. I recommend using distributed processing methods for efficient analysis of our large-scale temporal spatial data at the individual user level. For example, implementing Spark allows for parallel computing across multiple nodes, significantly reducing processing time and ensuring scalability for future data growth. This approach maximizes computational resources while minimizing processing time.

Moreover, refining clustering parameters involves experimenting with various algorithms, distance metrics, and parameter settings to optimize dwell estimation accuracy and robustness.

Finally, integrating additional data sources, such as transportation networks, presents an opportunity to enrich the analysis and gain deeper insights into spatial-temporal dynamics.

6. Conclusion

Dwell instances are observed to be concentrated in urban areas, transportation hubs, and commercial centers, reflecting typical locations where individuals spend time stationary, such as workplaces, residential areas, and recreational facilities.

The output of this assignment contains the identified individual dwells within a dwell location, the start and end time of the dwell, the duration of the dwell. The analysis provides valuable insights into user mobility patterns and spatial behavior, which on this basis, the actionable insights for urban planning, transportation optimization, and location-based services can be derived. Overall, this assignment has demonstrated the potential of GPS data analysis to inform decision-making processes and improve our understanding of human mobility patterns.