Assignment 2

Seyedeh Mina Mousavifar
CMPT826 - Data and Process Modeling and Analytics
University of Saskatchewan

February 29, 2020

Step 1. Preparation

This code is available in Step 1 section of *Assignment2.ipynb* file.

Step 2. Trip Definition

Trips Operationalization

For finding trips, first, I compare two consecutive rows for a user and discover whether it had changed its cell. Then, I gave id for these change cells to make them traceable as different trajectories. So, the stayed time in each cell can be calculated by grouping based on a specific user, grid and trajectory id. Afterwards, trips can be operationalized as staying in a cell more than a certain threshold will end the trip and leaving the cell start a trip. So we end a trip as soon as we know that it a user will stay in this cell more than certain threshold.
This code is available in Step2, Operationalizing trips section as trip_finder function.

Length of the trips (in grid cells)

This is the number of different grids seen until dwelling. So I grouped based on user and trip id in trip data and calculated the number of rows available in each group.
This code is available in Step2, Operationalizing trips section as trip_length function.

Number of trips

This is the number of trips founded for each user, which can be calculated by counting number of rows for each user in trip data or trip info data.
This code is available in Step2, Operationalizing trips section as trip_number function.

Time of each trip (in duty cycles)

The trip time is the summation of staying in each cell involved in a specific trip. However, if the user gets into a grid in which it will stay there more than N, I will immediately end the trip and count its staying time as one. Although I should note that for $N = 1$ trip time is the same as trip length because staying in a cell more than one duty cycle ends the trip.
This code is available in Step2, Operationalizing trips section as trip_time function.

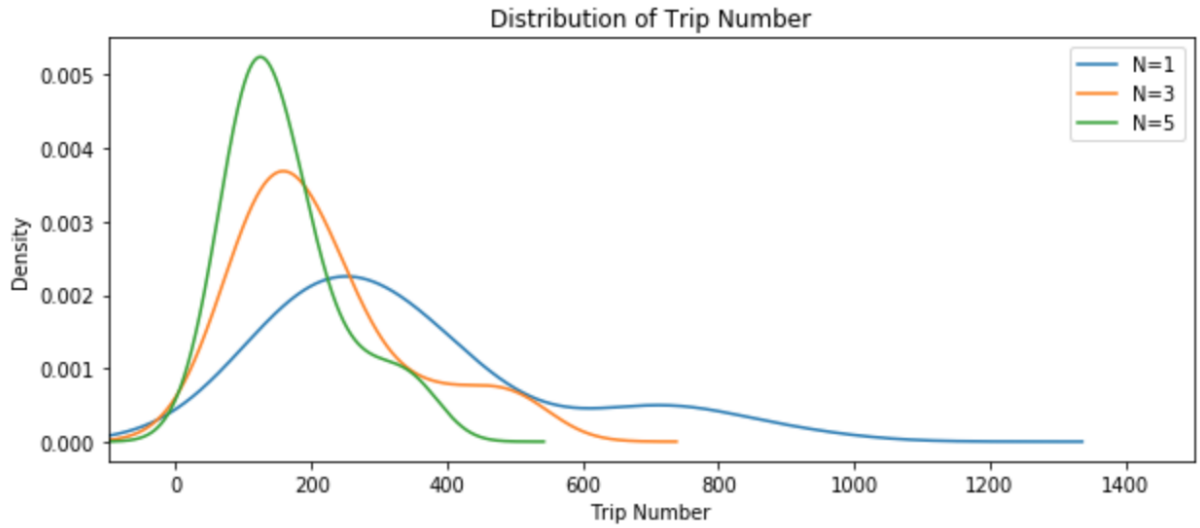


Figure 1: trip number distibution

Step 3. Presentation

Distributions for trip number

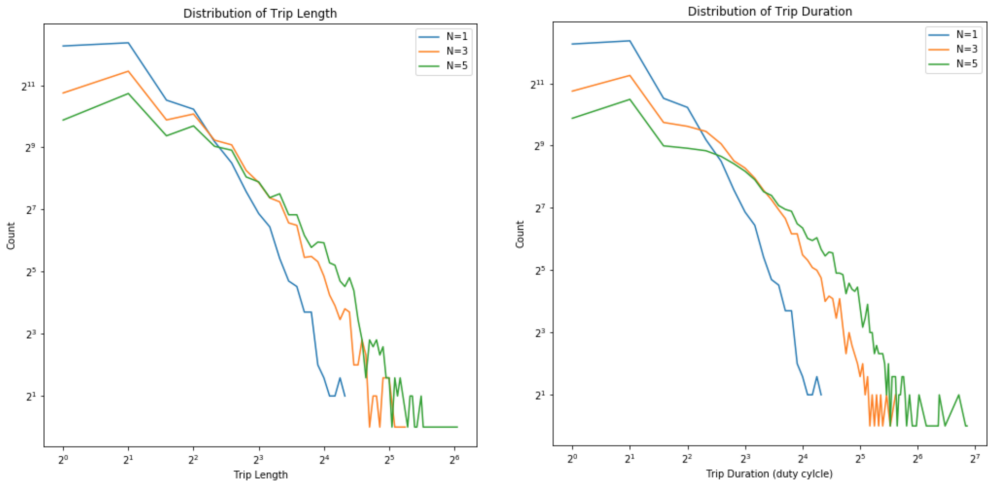
Because we have 41 records as the total trip number for each user, counting occurrences of different trip numbers would mainly result in one per trip number. Therefore, the plot having trip number as x-axis and count of it as y-axis wouldn't be informative enough. So, I plot the density of trip numbers, to show the distribution of trip number and find the concentration of trip numbers. Because our data has Gaussian distribution, linear axis fit it well.
Figure 1 shows distribution for trip number.

Distributions for trip length

For this plot, I first group by trip length and count occurrences of each trip length. When I use a linear axis, the curves had a steep fall, which didn't have much information. This shows a Power-law distribution. So I used the log-log axis, and after moving to the log-log plot, we can see patterns in our data.
Figure 2(a) shows distribution for trip length.

Distributions for trip duration

For this plot, I first group by trip time and count occurrences of each trip time. When I use a linear axis, the curves had a steep fall, which didn't have much information. This shows a Power-law distribution. So I used the log-log axis, and after moving to the log-log plot, we can see patterns in our data.
Figure 2(b) shows distribution for trip duration.



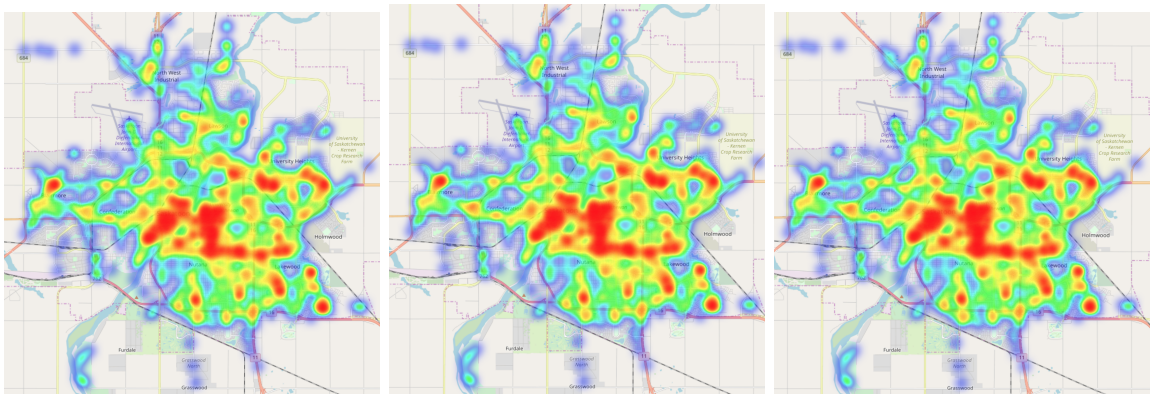
(a) trip length distibution (b) trip duration distibution

Figure 2: Distribution of trip length and trip duration

Heatmap over all participants which includes only trips

For this purpose, I select trips info and remove any dwell, but changing staying more than N to one, such that this grid can be seen as the destination of the trip. Then I calculated the summation of staying for each grid.
This code is available in Step3, Presentation section as trip_counter function.

For plotting heat-map, the grid labels were converted to x,y and then calculated its latitude and longitude with UTM. Furthermore, the summation of staying in a grid is used as grid heat. Figure 3 shows heatmap of trips for different N
This code is available in Step3, Presentation section as plot_heatmap function.



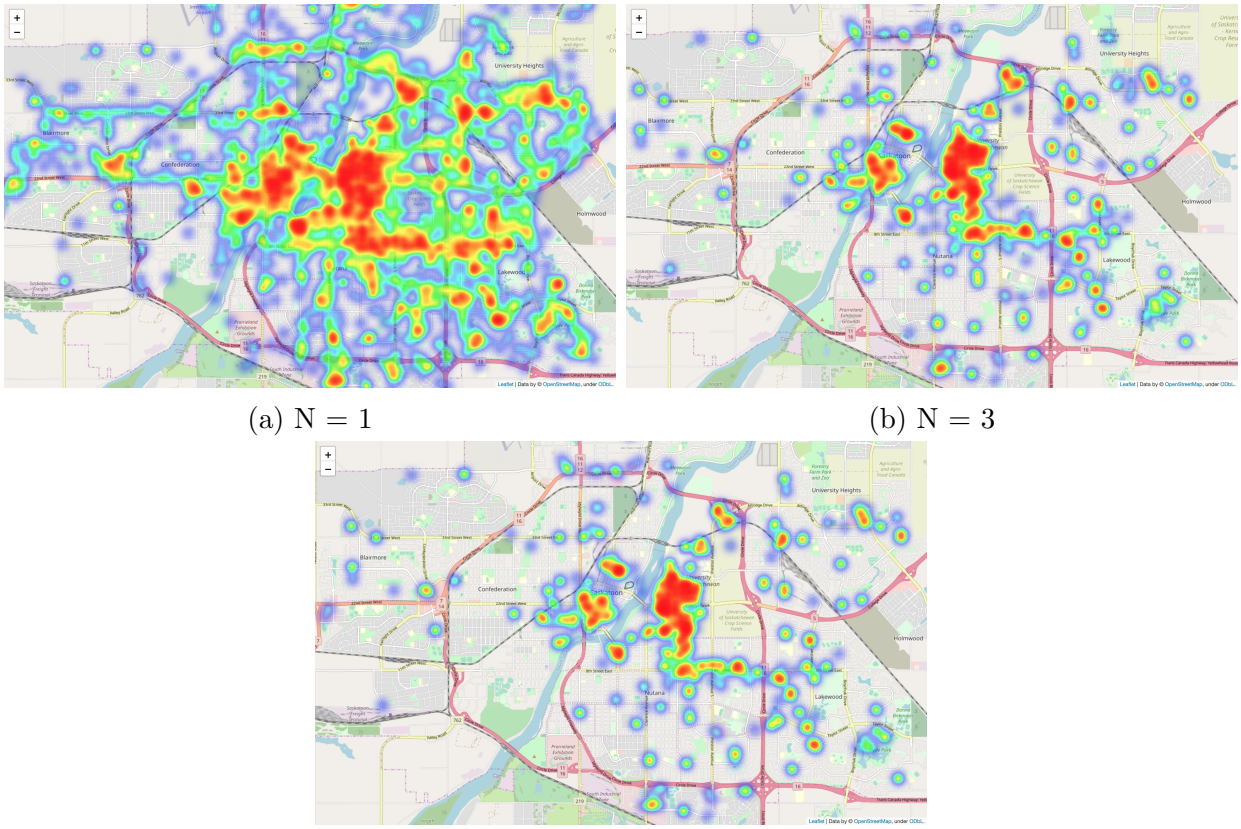
(a) N = 1 (b) N = 3 (c) N = 5

Figure 3: Heatmap of trips

Heatmap over all participants which includes only non-trips

For this purpose, I selected dwells which were records that have staying time more than N. Then I calculated the summation of staying for each grid. Figure 4 shows heatmap of non-trips for different N.

This code is available in Step3, Presentation section as nontrip_counter function.



(a) N = 1

(b) N = 3

(c) N = 5

Figure 4: Heatmap of non-trips

Step 4. Interpretation

Question 4.1

What is the impact of changing N on the distribution of trip number, length and duration?
As N increases, **Trip Number** decreases, **Trip Length** and **Trip Duration** increase. Because we are considering a higher threshold for stopping a trip, and consequently, we end trips later than before.

What trips are being captured and which are being ignored?
As N increases, we consider trips which are mostly for changing location and go to different destinations. Higher N, we will capture more meaningful dwells, such as going to work or getting back home. So by increasing N, we tend to capture more meaningful destinations. However, we lose the details of the trip.
For example, consider such trajectory; stay home for 30 duty cycles, then change cells and go to store for three duty cycles, then grab a coffee for two duty cycles, then changing cells, get to work and stay there for 14 duty cycles. With $N = 1$, each small stop will end our trip, but with a higher N, we will still capture the trajectory as one trip.

Given an example of a research question where the differences would be important, and an example of a question where they would be unimportant.
Consider choosing a hospital location in Saskatoon, the 8th street area. Our objective would be selecting a place that is accessible and has a short trip length and time. So, if we choose a small N, such as $N = 1$, we would capture lots of small trip length and time from lots of intersections. This might lead to choosing the hospital location in one of the intersections. However, this won't be an appropriate location because we have captured lots of insignificant dwells, such as staying behind a red light or shopping at a store as the starting point of the trip. In contrast, we know that the better approach would be finding trip length and duration from more accurate destinations such as home or work locations, which can be captured by a higher N.
Another research question can be choosing a coffee shop location in Saskatoon. Here, if we decide large N, we would eliminate lots of on-the-go places, and be biased to select our location near houses or work locations. However, the better approach is to choose a small N that can capture short trip lengths, and find trips that people only take to fulfil their needs and not to change their destination.
Nevertheless, consider choosing a supermarket location in Saskatoon. In this situation, it doesn't matter if we have long or short trip length and time, as long as we choose locations with high visit frequency, we would be able to select a suitable location.

Question 4.2

What distinguishing features did you see in the heatmaps?
For trip heatmap, we can see that most of the trips occur in university, downtown, and 8th street area. Moreover, as we increase N because we would omit less small dwellings, we have more red heats than before.
For non-trip heatmap, we can see that as we increase N, we can find destinations such as workplaces, homes more precisely because they can be seen as discrete points. Moreover, most of the non-trips occur in university, downtown. Because the participants were recruited from university and were students. Consequently, they spend most of the time at university and live near the university.
One important finding is that we see a massive decrease in heat in 8th street between trips heatmap and non-trips heatmap. This result is since 8th street includes many bus routes and has many stores and restaurants, which people seem to commute via this road and fulfill their needs but won't stay for a long time. Besides, the non-trip heatmap points are discrete, and many of them don't have any connection with each other, but in the trip heatmap, most of the grids are connected, which can be seen as a trip path.

Where there points included in either map (trip, not trip) at any N that seemed out of place? How would you change the operationalization to eliminate these points?
In figure 3, the trip heatmap, there are several points out of Saskatoon without any path showing a trip from Saskatoon to these areas. For filtering these points, we should find trips and then consider each grid label and check whether there is a $grid\_x \pm 1, grid\_y \pm 1$ grid label in this specific trip for this particular user or not. If the above condition doesn't hold, we should omit this record.
In figure 4, the non-trip heatmap, first, there is a dwelling located in the river, which should be removed by extracting corresponding grid labels and remove dwells in these locations. Furthermore, there are less strong points near highly concentrated points, e.g. near university heights, between McOrmond Drive and Attridge Drive, which is the result of GPS bouncing. For fixing this issue, we can add extra smoothing to our data by looking at the top 10% frequent locations, and if their nearby grid has a low dwell frequency, add to the related frequent grid and remove this nearby record.