CMPT 826 - Data and Process Modeling and Analytics

Assignment 1

This assignment is meant to provide you with practical experience in some of the low level filtering, aggregation and processing of geospatial data, as well as an introduction to the principle of operationalization. This assignment is focussed on geospatial data, but the overall principles and approach are common across many analytical tasks. You will be recreating a portion of the paper by Tuhin Paul et al (see readings folder) that describes the impact of changing spatial sampling size on common aggregate distributions used to describe human mobility. You will submit a report with sections answering questions or presenting results, as well as your Python code for the analysis. Please answer each set of questions in a single paragraph per step, with figures where appropriate or requested. For this assignment use the SHED10 database on crepe.usask.ca. Hand in both your Python scripts and a PDF document answering the questions below.

Step 1: Filter (10 Marks)

First, remove all participants who have returned less than 50% and 75% of total possible battery records.

Second, remove all GPS traces outside the city limits of Greater Saskatoon (52.058367, -106.7649138128), (52.214608, -106.52225318), and an accuracy of 100m or worse.

How many participants did you eliminate in each of the 50% and 75% thresholds? How many GPS records did you eliminate? Make a table outlining the number of GPS records for each filtering case. Plot all GPS for all participants as a heat map with and without sampled points. Are there any suspicious locations? How would you have filtered them if you had not filtered for Saskatoon?

Step 2: Stratify and Aggregate (20 Marks)

The next step is to specify the aggregation framework. Using the 50% or better participants and Saskatoon location data, aggregate by time by taking the average location every duty cycle, then transform data to UTM coordinates. To do this you will need to convert all the measured location values into UTM coordinates. I suggest using the proj library (pyproj for the Python wrapper) and the epsg code 32613. Once you have the locations in UTM coordinates you will be able to bin the locations according to the grid. Test this by generating a heat map at a 400 m grid size for all participants. What is the most commonly visited place by participants in Saskatoon? Describe two other commonly visited places based on your heatmap. Based on the heatmap name the top three neighbourhoods where participants live. Describe the method you used to infer home locations.

Step 3: Model Operationalization (10 Marks)

Based on the discussion in class, operationalize dwell time and visit frequency. In your document, note exactly the algorithms you used for each including any parameter choices you might have made. Comment the code thoroughly. Have mercy on poor Luana, who will be marking your code.

Step 4: Model Interpretation (20 Marks)

Using steps 2 and 3, calculate the distributions of dwell time and visit frequency per participant at 100, 400 and 1600 m grid sizes, on a per participant basis. Plot heatmaps for each resolution aggregated across all participants. Plot dwell time and visit frequency aggregated over all participants at each of these distributions as different curves in a single graph. What are the properties of these curves (what distributions are you likely looking at)? How has the shape properties changed with spatial resolution?

Code Quality (10 Marks)

Code Comment Quality (10 Marks)

Note 1: Ethica Data uses a one minute on, four minutes off duty cycle, hour aligned. However, it also allows some leeway for the operating system to shift the exact start time to save energy by batching Ethica's requests with other tasks. If a duty cycle has been triggered on a five minute boundary, then the next duty cycle is allowed to be scheduled at the discretion of the OS. If the OS schedules that duty cycle at a time other than at a five minute boundary, the next duty cycle is forced to be on a five minute boundary. You will have to take this into account when aggregating over duty cycles.