Multi-visit inference of mobility data

Amir Mohammad Esmaieeli Sikaroudi, amesmaieeli@arizona.edu

University of Arizona, CSC696

October 12, 2022

Problem definition

To learn mobility patterns of cities, mobile tracking is a famous approach to gather large-scale data. However, the data needs to be anonymized for privacy considerations. Commonly three changes are applied to mobile mobility datasets:

- Trajectories are not represented from individuals but aggregated for each Point of Interest (POI)
- Sequence of visits of each individual is removed
- Removing trajectories with very few number of trips from a source region

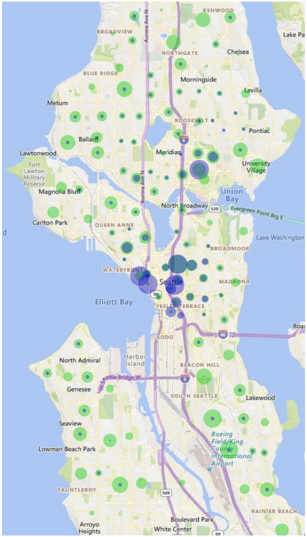
In this project it is tried to infer the sequence of visits from individuals perspective. Individuals are generated based on the demography of a city and the important question is that on the average what sequences of visits are more likely to happen from different geographical parts of the city.

What data is available

SafeGraph mobility dataset:

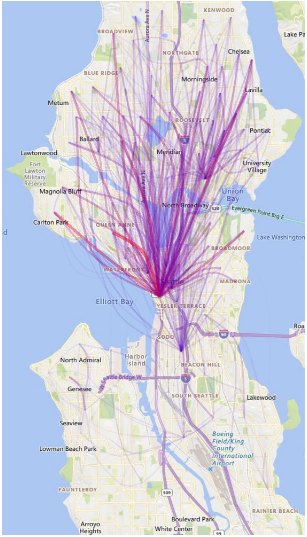
- POI's latitude and longitude
- POI's North American Industry Classification System (NAICS) code which represent the type of the POI
- Number of visits per month to a POI
- Number of unique visits per month to a POI (no revisit)
- Frequency of visits to a POI for each day of month
- Frequency of visits to a POI for each day of week (aggregated for 4 weeks)
- Frequency of visits to a POI for each hour of day (aggregated for 30 days of month)
- The frequencies of rips from different source Census Block Groups (CBGs) (for a month)
- The median travel distance to a POI (for a month)
- The frequencies of stay duration ranges at a POI (for a month)

What data is available



Source: Graphical Models of Pandemics (2021), derived from US Census bureau

What data is available



Source: Graphical Models of Pandemics (2021), derived from SafeGraph

What data is available

Federal Information Processing Standard (FIPS):

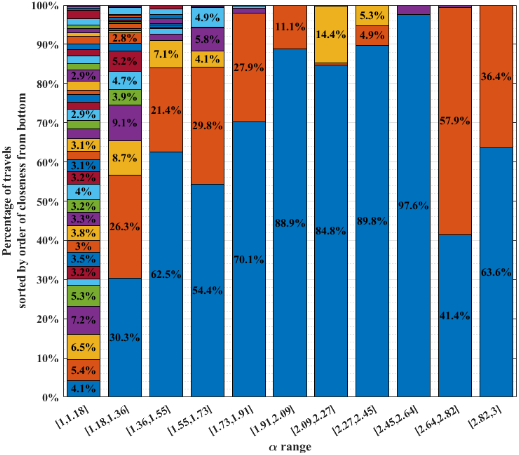
- The geographical polygon of CBGs with latitude and longitude
- Population density of each CBG
- Age range frequencies of each CBG

The needs for different age groups:

- The general needs for different age groups (Facs: a geospatial agent-based simulator for analysing covid-19 spread and public health measures on local regions (2022))
- The needs are manually mapped into NAICS codes

What data is available

The preference of visiting nearby POIs based on the distances:



Source: Derived from analyzing the SafeGraph for entire USA

Proposed method

A LDA like model is proposed to infer the sequence of visits.
Since it can be hard to jump into plate notation, generative process is discussed:

- For each visit
  - Sample a POI
  - Sample a day of month distribution
  - Sample a day of week distribution
  - Sample an hour of day distribution
  - Sample a duration range distribution
- For each individual
  - Sample a source CBG
  - Sample the age group
  - Sample a need
  - For each sequence (fixed number of sequences, is infinite sequence possible?)
    - Sample a visit based on distance (?)
    - Sample a day of month distribution
    - Sample a day of week distribution
    - Sample an hour of day distribution
    - Sample a duration range distribution

Previous works

The most relevant works to the proposed project are related to inference of intention of trips which used graphical models:

- Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model (2020)
- Individual Mobility Prediction in Mass Transit Systems Using Smart Card Data: An Interpretable Activity-Based Hidden Markov Approach (2021)
- Where Did You Go: Personalized Annotation of Mobility Records (2016)

There are also deep learning models for forecasting the sequence of trips. The key point is that the sequential data is available in their works.

- DeepMove: Predicting Human Mobility with Attentional Recurrent Networks (2018)

Validation

The problem is related to unsupervised learning. Perplexity is the first measure that come to mind.

The difficulty is that the inferred sequence can not be assessed with predictive distribution because the priors are different from posteriors (any ideas?)

One approach of validation is to build the model on a city smaller/larger and test it on another larger/smaller city. The data for entire US for cities above 85K population is available. The total number of visits on a test city should be close to what the dataset provides.

Maybe the model can be represented by a Recurrent Neural Network, however, the data structure doesn't seem to fit (any ideas?).