Spatio-Temporal Prediction of Human Mobility

The 2021 Aspiring Scientists' Summer Internship Program

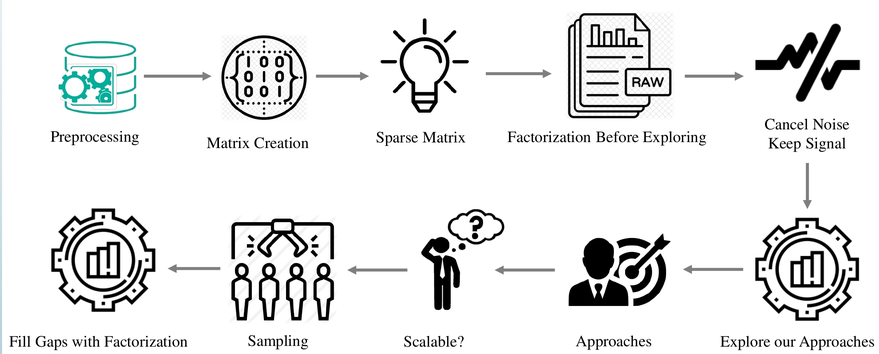
Samiul Islam[1], Dhruv Gandhi[1], DEREK GOH[2], Taylor Anderson[3], Amira Roess[4], Tim Leslie[3], Hamdi Kavak[1], Andreas Züfle[3]

1. Computational and Data Sciences (CDS)
2. Aspiring Scientists Summer Internship Program, College of Science (ASSIP)
3. Geography and Geoinformation Sciences (GGS)
4. Department of Global and Community Health (GCH)

Abstract

There have been many studies on location prediction that have tried to predict the number of visitors at a certain location; however, these studies suffer from some limitations due to the lack of data. This study attempts to create a location prediction system based on data for Fairfax County from SafeGraph. SafeGraph data provides us with a sample of approximately 10 percent of the number of visitors from each census block group to each place of interest from Jan 2018 to Jun 2021. Tensor factorization was used to cancel out the noise that is present in the SafeGraph data. We developed three baseline prediction models to compare against more sophisticated approaches: weekly rolling average [model 1], using previous 4 weeks [model 2], and using previous 4 weeks weighted [model 3]. Other approaches used were long short-term memory (LSTM), linear and polynomial regression,. The baseline approaches, specifically model 2 and model 3, have consistently performed better than the more sophisticated approaches and tensor factorization has always influenced our models positively. Given that we have access to this abundance of data, using the power of factorization and simple baseline algorithms, businesses and entities can predict human mobility and identify a potential group for their mutual betterment through marketing and advertisements.

Workflow



Approaches

- Previous week's data point (Model 1)
- Average of last one month (Model 2)
- Weighted average of last one month (Model 3)
- Regression
- Long Short-Term Memory (LSTM)

Conclusion

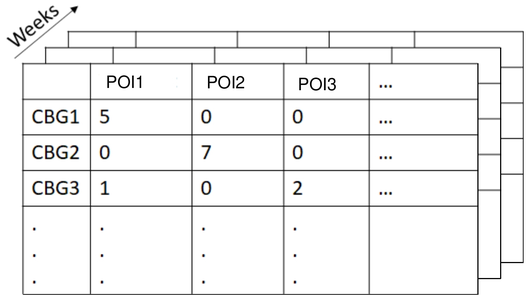
The initial results of our research are promising and lead us to believe that the baseline approaches are better than the more sophisticated approaches we have tried. Tensor factorization seems to aid in the accuracy of all the models. We have only applied our models on data from Fairfax County, but we would like to test our models on a larger geographical subset of the United States and create more models to compare to our baseline models.

Future Work

- Finding the key-trends between CBGs and POIs and see if they sustain over a longer period and/or different geographical areas
- Develop and implement efficient (possibly multi-threaded) approaches to predict mobility for a larger area
- Producing more sophisticated approaches and compare those against the baselines we have so far such as:
  - Croston
  - Poisson Regression
  - Auto Regressive Integrated Moving Average (ARIMA)
  - Exponential Smoothing

Introduction

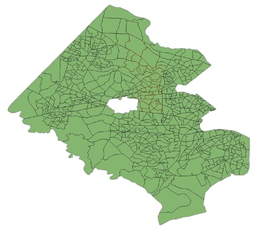
Many of us have smartphones with apps that connect to the cloud, which then collect data on our location information. SafeGraph, a data company, collects this data and then distributes it for all kinds of uses ranging from demographic analysis to advertising. This data enables us to create human mobility models to predict the number of visitors to a place of interest (POI) from a census block group (CBG). A place of interest is any location where people gather, such as a gas station, grocery store, or mall. Census block groups are the smallest geographical units for which the United States Census Bureau publishes sample data. SafeGraph provides us with weekly data from January 2018 to June 2021 for about 10 percent of all POI visits from a CBG. We are using a subset of that data which is restricted to Fairfax County, where we have 649 CBGs and 15197 POIs. Our objective in this paper was to create an algorithm that predicts the number of visitors from a CBG to a POI in the upcoming weeks.



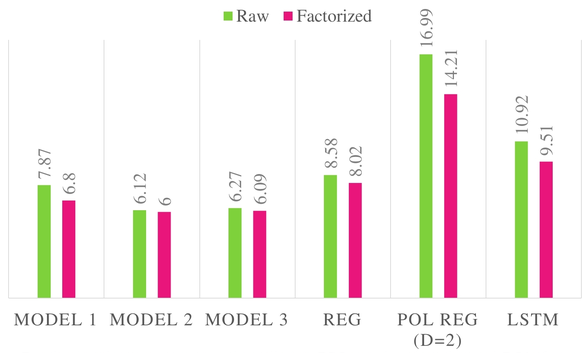
	POI1	POI2	POI3	...
CBG1	5	0	0	...
CBG2	0	7	0	...
CBG3	1	0	2	...
.	.	.	.	
.	.	.	.	
.	.	.	.	

Results (RMSE Comparison)

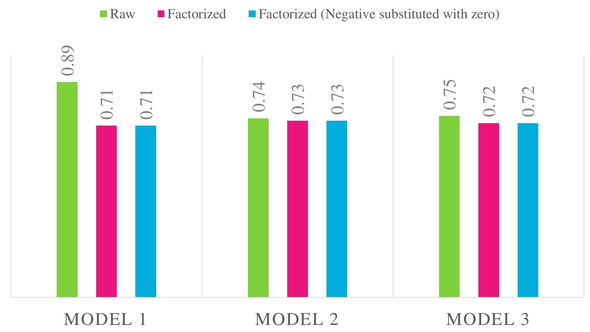
- Randomly selected CBG
- 15197 POIs (All the POIs)
- 104 of 182 weeks
- Dec 31, 2018, to Dec 28, 2020
- 1,580,488 data points



- 649 CBGs (Entire County)
- 1670 of 15197 POIs
- 52 of 182 weeks
- Jan 1, 2018, to Dec 30, 2018
- 56,359,160 data points



Predicting how many people will visit 500 randomly selected POIs from a randomly selected CBG



Predicting each CBG-POI pair for those most visited POIs

Acknowledgements

- This work is supported by National Science Foundation Grant #2109647 titled "Data-Driven Modeling to Improve Understanding of Human Behavior, Mobility, and Disease Spread"
- This work is supported by the George Mason University Aspiring Scientists Summer Internship Program (ASSIP)
- This work is supported by a 2021 Summer Team Impact Grant of the George Mason University Office of the Provost and Executive Vice President

Data Source: https://docs.safegraph.com/docs/weekly-patterns