Project Proposal – CP255
Juan Caicedo
SID: 3033084329

Reveling and making sense of mobility patterns of public transportation users: Case study of Bogota's BRT system - Transmilenio.

Introduction

Since the implementation of smartcards to speed and have better control over the fare collection systems, public transit agencies have been collecting an unprecedented amount of data. Most of the smartcards have a unique identifier that makes it possible to track how users use their cards while they interact with public transportation systems. Although this information is rich in the longitudinal sense, it lacks the standard variables that are used in transportation demand models, such as trip characteristics and demographics. For instances, some system only records tap-in location and time, but no information is given on the route the user takes, alighting stations, trip purpose, income, age, or gender. This proposal aims to infer additional information for the users by understanding travel patterns and also by combining household travel surveys and census data to try to match observations that may have similar behavior.

Literature Review

Since smart card data collection is passive, different data mining methodologies have been used to reveal travel characteristics of the users, such as destination estimation and home and work location. For many systems, where only boarding validation is required, estimating the destination is relevant to construct origin-destination (OD) matrixes. The assumption here is that the destination of a trip is the origin of the next trip, and the destination of the last trip is home. One of the problems with this estimation is that it required at least two trips to infer destination. Additionally, it is unknown to the researcher if an individual made other trips in different transportation modes (Kurauchi & Schmocker, 2017). Trepanier, Tranchant , & Chapleau, 2007 improve this estimation by considering previous days of travel, with the idea that previous transit habits are relevant to infer today's destination. Ma and Wang (2014) used smartcard data and GPS data from the buses to estimate the origin and destination of the flat-fare buses in Beijing, China. They use a Bayesian tree to calculate the stations' probability based on the same individual historical data. Muniziaga and Palma (2012) added a minimum bus time and walk to transfer time to infer the destination stop of a transit transfer in Santiago de Chile. More recently, research in OD estimation focuses on linking the transit OD matrix with the Transportation Analysis Zones OD matrix (Hussain, Bhaskar, & Chung , 2019).

Home and work locations have been inferred from both CDR records and smart card data. These papers use a station rank classification for each individual, where rank one is the most visited station and so forth. Similar to household travel results, most visited locations are correlated with home location. Second, most visited locations tend to be mandatory trips locations such as

work and education; other ranks could be classified as other activities (Samiul, Scheneider, Ukkusuri , & Gonzalez , 2012). Zou et al. (2018) use one-week data form the Beijing underground system to estimate home location using a center point-based detection algorithm, in which the average first and last station in a day trip is located close to home. This assumes that the first and last stations of a daily trip are highly correlated with home location.

To the author's knowledge, household travel surveys, and census data has not been combined with smartcard data to complement each other's.

Data sources

For this project, we have available three different sources of data.

- Transmilenio data: Contains anonymized smartcard ID and transactions from 03/12/2018 to the present day. Each observation contains the boarding station, boarding time, unique ID for each card, credit previous the transaction, fare price, and type of card. The system records about 2 million transactions per weekday and 0.5 million transactions per weekend days.

- Household travel surveys (HTS): In Bogota, collects household travel surveys every four years. However, respondents among surveys are not the same as there are different randomization processes are generated for each year's survey. For this research project, we will use HTV 2015 and HTV 2019, as they share a similar timeframe with the smartcard data.
  - HTS 2015 is publicly available at the Secretary of Mobility website. It contains travel diaries for one day from ~28000 households. There are four main tables: (1) Household table, (2) person table, (3) Trips table, and (4) Trip legs table.
  - HTS 2019 is not yet publicly available. However, the Bogota Secretary of mobility will do so by the end of February 2020. The survey tables will be available on their website.

- Census 2018: This data is public trough the Colombian National Department of Statistics. I have not explored the information available here, but I assume it contains information and demographics about the general population, which may be similar to the US census data.

Processing and data analysis

As the universe of data is within the definition of Bigdata (more than 2 million observations per day, for almost two years). I will simplify my analysis to a random sample of 50.000 users. A random sample not only generates a representative sample of the population but also is easier to manage and faster to run calculations and plots. Eventually, the plan is to use the entire dataset.

A data cleaning process is necessary to remove observations that were created for test purposes, or those that may have weird values, such as tap-in times at 1 am when the system is closed.

I will treat each unique card ID as the unit of analysis. As there are different tap-in times for different smartcard data, I will represent each observation with a 3D matrix, representing the time of boarding, day of boarding, and station of boarding, respectively. To discretize the time variable, we will divide it into chunks of 30 minutes. The matrix will contain a one if a smartcard records a boarding time within those 30 minutes, in a specific day at a given station, and zero otherwise. As a visualization tool, I will plot two dimensions of a given observation. The purpose of this is to visualize what are the travel patterns for an individual, for instance, I expect to see that a worker consistently uses two stations throughout the day and is pretty consistent in time. I could assume that the station that uses in the morning corresponds to the station closest to the home location, and the station used in the afternoon is closest to the work location.

For the household travel surveys, I will first do a data exploration to understand the population and the variables in the tables. I will use this information to determine "catchment areas" for each of the Transmilenio stations. For each station, I will try to understand the demographics and the purpose of the trips of respondents that reported using Transmilenio in any leg of their trips. I will try to compare 2015 to 2019 data and understand if there is a time-stability of the demographic characteristics for each catchment area. Census data will be used to estimate more representative demographics for catchment areas.

Problems

There is the risk that I may not have full access to the HTS 2019, as the estimate for getting that data obtained through an informal conversation with the public agency. Also, I am not familiar with what data will be public and may not be well-documented. There is also the risk that HTS for the two years may not be comparable, as the population is not constant across surveys, and both surveys used we conducted by different consulting companies in opposing governments.

The Transmilenio smartcard data is not open data. This means that I need to be extra careful not to disclose the information to anyone.

Probably, a match of the smartcard data and the HTS is not clear or not possible to find because both datasets are anonymized. Additionally, smartcard data contains longitudinal travels, while HST may only contain one day of travel diaries. That is, I will not have the ground truth to compare my results. However, I will try to use statistical methods to try to match similar behaviors. Also, I would need to assume that each smartcard data is used by a singular user only.

Bibliography

Hussain, E., Bhaskar, A., & Chung , E. (2019). Use of smart card data for zonal level public transit OD matrix estimation: literature review and research gaps [Working Paper]. *(Unplublished)*.

Kurauchi, F., & Schmocker, J.-D. (2017). *Public Transport Planning with Smart Card Data.* Boca Raton, FL: CRC Press.

Ma, X., & Wang, Y. (2014). Development of a Data-Driven Platform for Transit Performance Measures Using Smart Card and GPS Data. *Journal of Transportation Engineering, 140*(12), 04014063.

Muniziaga, M., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies, 24*, 9-18.

Samiul, H., Scheneider, C., Ukkusuri , S., & Gonzalez , M. (2012). Spatiotemporal Patterns of Urban Human Mobility. *J Stat Phys*(151), 304-318.

Trepanier, M., Tranchant , N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems, 11*(1), 1-14.

Zou, Q., Yao, X., Zhao, P., Wei, H., & Ren, H. (2018). Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. *Transportation, 45*(3), 919-944.