Transportation Affordability, Accessibility, and Reliability Matrix:

Data Science Technical Plan

Contacts: Joe Hellerstein and Ryan Maas

May 1, 2018

This document discusses technical planning for the DSSG 2018 summer project [Transportation Affordability, Accessibility, and Reliability Matrix](https://drive.google.com/drive/folders/1kFpgp--kswQkiXdZN6xGTLRWpZqoyJ9N). We have renamed the project to “Seattle Mobility Index Project”

# First Deliverable

Based on our current knowledge of the available data, the first deliverable will be software that calculates estimated relative costs of reaching the basket of destinations as a function of the least and most expensive modes by block group for a market basket of destinations. We refer to this as the **Affordability Calculator**.

SB: I think we will need to build the **Mode Choice Calculator** before the Affordability Calculator. Although a rather simple calculation, the Mode Choice Calculator provides the relative quantity of modes available to reach the basket of travel destinations, within designed parameters.

SB: Also, the Basket of Destinations Calculator is an integral and very interesting part of the project, although at the request of DSSG we are doing a lot of this analysis before the summer. The first deliverables could include refining the Basket of Destinations Calculator and incorporating it into the workflow.

Instead of using a complicated, multi-step transportation land use model (IE Oregon DOT model), the Basket of Destinations Calculator uses a low-cost, repeatable, and scalable algorithm that creates a list of actual destinations for every origin. The universe of possible destinations is a combination of a curated list of citywide locations (GoogleMatrix\_Places\_Citywide.csv) and a list of local destinations pulled from the google places api (GoogleMatrix\_Places.csv) with google-place-search.ipynb.

To evaluate the basket algorithm, I am proposing a new feature called the "proximity ratio," which is the number of trips less than 2 miles vs the number of trips between 2 and 10 miles. Trips over 10 miles are outside the city and the volume shouldn’t really vary based on the origin. The python notebook has a function to compare the generated basket with calculated values from the PSRC.

This mechanics of this algorithm can be visualized here:

https://public.tableau.com/views/Basket\_of\_Destinations/Dashboard?:embed=y&:display\_count=yes

The Mode Choice Calculator and Affordability Calculator takes as input the **market basket description** (blockgroup\_basket.csv, to be provided by the Project Leads). We propose a tabular structure for the first version of the market basket description. The columns are:

* origin(blockgroup)
* destination coordinate (or latitude, longitude)
* destination key

Using the market basket description, the City will access the Google Distance Matrix API and generate a dataset, distance\_matrix.csv, that will consist of travel times, distances, and modes between each block group and its basket of destinations. Walk and bike data will be collected once. Transit will be collected for a 14-hour period for both weekdays and weekends. Drive data will be collected on an ongoing basis for the duration of the project. The columns are:

* origin(blockgroup)
* destination key
* travel mode
* distance (miles)
* travel time (minutes)
* datetime

# Data Sharing

To facilitate sharing and analysis, we will write software that automatically creates a SQL file named transportation.db file from the CSV files provided by the Project Leads. The DSSG team will work with the SQL file, not the CSV files; access to the SQL file will be provided using a technology such as [quilt](https://quiltdata.com/).

The first version of transportation.db will be built from distance\_matrix.csv.csv, blockgroup\_basket.csv, and GoogleMatrix\_Places\_Full.csv. The table names will be the file name; the columns will be the columns in the CSV files, with some small modifications that improve usability.

# Questions

1. What is the relationship between block group, zipcode, and neighborhood (uv\_origin in df\_Trip\_Processed\_Agg\_Village\_Origin.csv)? How do we join data that use these different area specifications? This is for reference and we don’t need this for calculations. I have a custom blockgroup correlation file for crosswalking block groups to other common geographies. This is for reference, for example, because most people don’t know there way around by blockgroup number. The basis of the correlation file is here: http://www.seattle.gov/Documents/Departments/OPCD/Demographics/GeographicFilesandMaps/SeattleCensusBlocksandNeighborhoodCorrelationFile.xlsx
2. Are the units of duration minutes? yes
3. How do you plan to obtain data on transportation times for walking, bus, and bike (or other personal transport)? We will download the data via the google API. Walk, bike, and transit can be done prior to the project start. If we have time and decide to incorporate other modes such as bikeshare, rideshare and carshare, I can probably come up with some sort of density function for each blockgroup and we can make some general assumptions on availability based on density.
4. How do we estimate the cost of a trip made by walking? by bike? We will look at time component costs and national standard calculations and informed by discussions with domain staff and transportation modelers.
5. Is it acceptable to use a fixed cost-per-mile to calculate the cost of driving? Yes, we would propose using the national standards for costs per mile. It’s my preference but there are other ways, too.
6. We’d like to commit to a format for each CSV file (although we can agree to change the format at a later time). This means that each CSV file has specified heading names for columns and specified data types for column values. Great, if you can tell me the best way to format them, we can put everything so far in that format. Same for file names and code variables, etc. I’d like to keep everything oracle and sql compliant for an easy transition to production.
7. When can we get a first version of the market basket description? We should be able to provide a version in the next two weeks- see attached draft basket-generation-calculator.ipynb and the tableau viz.
8. How are the aggregates in df\_Trip\_Processed\_Agg\_Village\_Origin.csv computed from the PSRC survey data [here](https://www.psrc.org/household-travel-survey-program)? I used python scripts to do, among other things, geocoding, aggregation and pre-processing of the data.