U.S. Household Commuting Dataset and Transportation Fairness

Hao Hao[1], Hai Wang[2], and Peter Zhang[1]

[1]Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]School of Computing and Information Systems, Singapore Management University, Singapore

Abstract

Efficient and fair transportation planning creates opportunities and equity for jobs, health care, and education. Therefore, data consolidation for transportation systems provides basis for evidence based policies. In this work, we construct a dataset that documents home-to-job commuting time and distance information for the 100 most populated U.S. urban areas. Our dataset builds on the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Dataset [U.S. Census Bureau, 2018b], which provides origin (home) and destination (job) location information for households. For these origin-destination (OD) pairs, we derive commuting time and distance information under different travel modes, each a combination of walking, public transit, and ridesourcing. We construct data under different modes so policymakers and researchers have information about alternatives and can perform what-if analysis. Towards the end of this data sheet, we document a sample use case to illustrate this goal.

1 Dataset Description

The dataset is available for download at this link `https://drive.google.com/drive/folders/1AacvpvQqh-boCpA9Q_Tvadfl5mj_L6NC?usp=sharing`. Each instance in our dataset contains commuting time and distance information for one OD pair. We randomly sampled 100,000 OD pairs from the U.S. Census Longitudinal Employer-Household Dynamics Dataset [U.S. Census Bureau, 2018b] (1,000 data points each for the 100 most populated urban areas, the list or urban areas with the corresponding indices can be found under `metadata/Urban_Area_ID.csv`). In particular, each instance contains travel time and distance information for five different travel modes, as documented in Table 1. The definition of columns of the dataset is provided in Table 2, with sample instances of the dataset shown in Table 3. The data files under Google Drive is organized as follows: `data/Duration_M1M5_{Urban_Area_Index=i}.csv` and `data/Distance_M1M5_{Urban_Area_Index=i}.csv` each

Table 1: Description of five travel modes. FM = first-mile, LM = last-mile. For the instances that do not have public transit available between OD pairs, only Mode Five (ridesourcing or driving throughout) travel time/distance are provided.

Travel Mode	Origin to FM Stop	FM Stop to LM Stop	LM Stop to Destination
Mode One	Walking	Public Transit	Walking
Mode Two	Walking	Public Transit	Ridesourcing
Mode Three	Ridesourcing	Public Transit	Walking
Mode Four	Ridesourcing	Public Transit	Ridesourcing
Mode Five	Ridesourcing from Origin to Destination		

Table 2: Definitions of the columns. For each urban area, there are two data tables, one for travel time, and one for travel distance.

Column	**Definition**
OD Pair	Index of the OD pair
FM Walk	Duration (seconds)/distance (meters) of first-mile by walking
LM Walk	Duration (seconds)/distance (meters) of last-mile by walking
FM Drive	Duration (seconds)/distance (meters) of first-mile by ridesourcing
LM Drive	Duration (seconds)/distance (meters) of last-mile by ridesourcing
M1	Total duration (seconds)/distance (meters) by Mode One
M2	Total duration (seconds)/distance (meters) by Mode Two
M3	Total duration (seconds)/distance (meters) by Mode Three
M4	Total duration (seconds)/distance (meters) by Mode Four
M5	Total duration (seconds)/distance (meters) by Mode Five

contains the travel time and travel duration information of $1,000$ OD pairs within the ith urban area.

Table 3: Sample instances of the travel duration data table.

OD Pair	FM Walk	LM Walk	FM Drive	LM Drive	Transit	M1	M2	M3	M4	M5
0	810	367	190	177	4,643	5,820	5,630	5,200	5,010	1,433
1	714	134	125	46	3,639	4,487	4,399	3,898	3,810	1,086
2	643	144	172	118	480	1,267	1,241	796	770	362
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

2 Dataset Construction

The following steps have been performed to collect and process the data:

1. **OD commuting trip information:** The block-group level OD pairs for U.S. workers' commuting trips are obtained from Origin-Destination Employment Statistics published by US Census Bureau [U.S. Census Bureau, 2018b]. This original dataset contains a total of 621,011,910 OD pairs collected for 53 US States/Territories. Each OD pair is characterized by the Residence Census Block Code and Workplace Census Block Code. Each Census Block Code is then mapped to the (latitude, longitude) coordinates of its block internal point. The internal point is not a centroid and the only guarantee is that it is inside the block.

2. **Sampling OD pairs for each urban area:** The shape files of the top 100 most populated US urban areas are obtained from 2018 US Census [U.S. Census Bureau, 2018a]. For each urban area, 1,000 OD pairs that have both the origin and destination within the urban area are sampled uniform randomly. As an example, the OD pairs for the top 4 most populated US urban areas are visualized on maps in Figure 1.

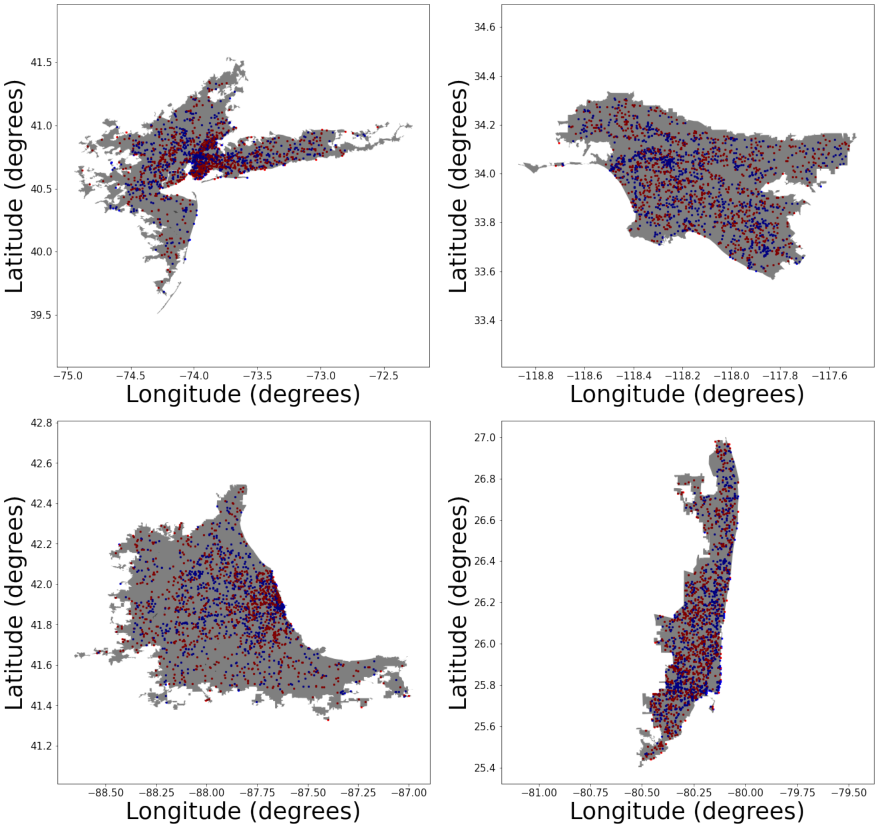

Figure 1: OD pairs sampled for the top 4 most populated US urban area. Red nodes denote the origins, blue nodes denote destinations. Top-left: New York–Newark, NY–NJ–CT. Top-right: Los Angeles–Long Beach–Anaheim, CA. Bottom-left: Chicago, IL–IN. Bottom-right: Miami, FL.

3. **Generating travel time and distance of five modes for each OD pair:** For each OD pair, we first obtain the Mode One travel information via Google Map API [Google Maps Platform, 2021]. The total OD travel time/distance of Mode One is obtained along with the travel time/distance of its three segments: first mile walking; public transit; last mile walking. In addition, the coordinates of the first-mile and last-mile transit stops are extracted. The ridesourcing travel time/distance of the first/last-mile is again generated using Google Map API for travel Modes Two, Three and Four. Finally, the travel time/distance for Mode Five is generated by querying OD travel direction in driving mode directly. All travel information are queried with a departure time of 8:00 AM local time, reflecting the traffic condition during the morning peak.

Data Limitations and Cost. Given the Google Map API query cost and time constraints, a dataset of 100,000 OD pairs were created from the original data source of 621 million OD pairs. In addition, we only sampled OD pairs with both its origin and destination within an urban area, this may induce bias as urban areas typical enjoy better access to public transit. The overall cost for creating our dataset is roughly $2,000 U.S. dollars.

3 Dataset Summary

3.1 Summary Statistics

The summary statistics of travel duration/distance across the five travel modes are shown in Tables 4 and 5. The summary statistics for the percentage of OD-pairs with access to public transit within each of the 100 urban areas is documented in Table 6.

Table 4: Summary statistics (min, max, mean, and standard deviation) of travel duration (in seconds) for five travel modes of 100,000 OD pairs.

Travel Mode	**Min.**	**Max.**	**Mean**	**Stdev.**
Mode One	174	21,570	4,421	2,570
Mode Two	169	21,284	4,047	2,504
Mode Three	184	21,356	3,935	2,489
Mode Four	122	21,070	3,561	2,422
Mode Five	30	6,632	1,098	581

Table 5: Summary statistics (min, max, mean, and standard deviation) of travel distance (in meters) for the five travel modes of 100,000 OD pairs.

Travel Mode	**Min.**	**Max.**	**Mean**	**Stdev.**
Mode One	428	239,520	21,519	16,683
Mode Two	481	239,549	21,769	16,720
Mode Three	448	239,520	21,691	16,708
Mode Four	526	239,549	21,941	16,745
Mode Five	137	149,516	17,770	13,527

Table 6: Summary statistics (min, max, mean, and standard deviation) of the percentage of OD pairs with access to public transit over 100 urban areas.

Min.	**Max.**	**Mean**	**Stdev.**
0.4%	92.8%	59.9%	20.1%

3.2 Histogram

Figure 2 shows the histograms of OD travel duration and distance under five different travel modes.

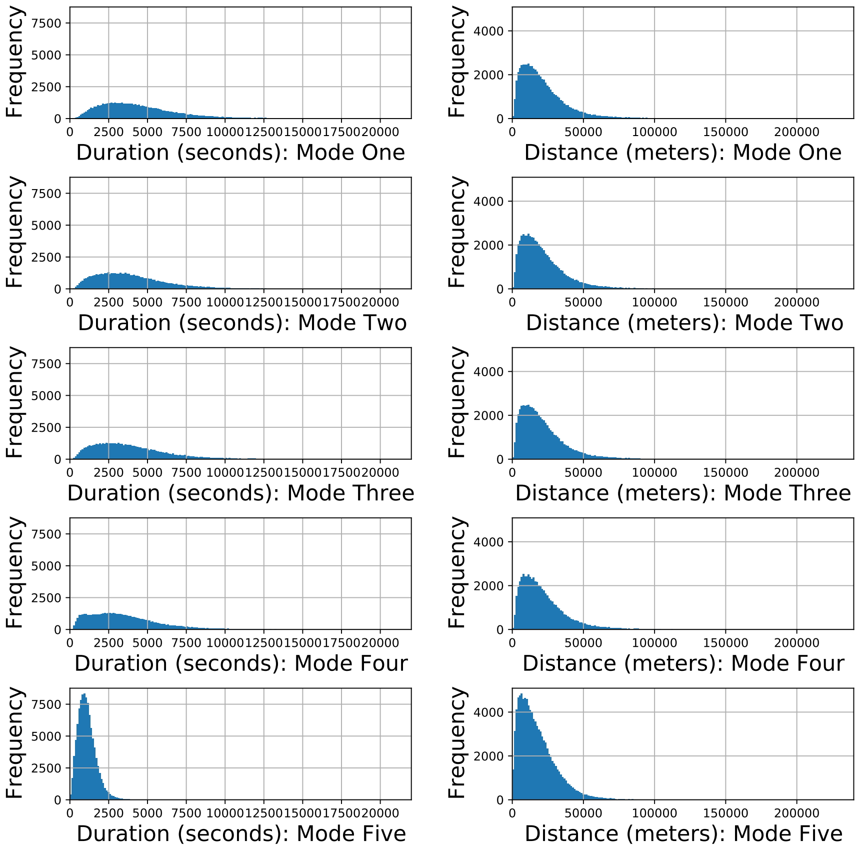

Figure 2: Histograms of OD travel duration and distance under five travel modes.

4 Dataset Use Case: Does First/Last Mile Ridesourcing Reduce or Increase Transportation Equity?

The design of first/last mile transportation systems (F/LMTS) has been one of the foci in recent transportation research [Wang and Odoni, 2016, Wang, 2019]. But practically, some questions remain: Can these new systems actually provide travel benefit? How much benefit is there? Is it equally beneficial for every neighborhood regardless of their income level?

A recent study [Reck and Axhausen, 2020] argued that low income families may have lower adoption rate for first and last mile transportation services, suggesting the potential widening of employment inequality gap due to advances in transportation systems. In this vein, we provide a use case of our dataset, investigating the relationship between the household income level and the utility gained from F/LMTS. We support this use case with the addition of block group level income data obtained from [U.S. Census Bureau, 2019].

1. **Travel Time v.s. Income Level (Figure 3).** We perform this analysis with the addition of block group level income data obtained from [U.S. Census Bureau, 2019]. The relationship between income level and commuting time for a travel mode can be plotted as shown in Figure 3. Each point in the scatter plot represents one OD pair. A trend with negative slope may suggest an uneven distribution of public transit

resources for rich and poor neighborhoods.

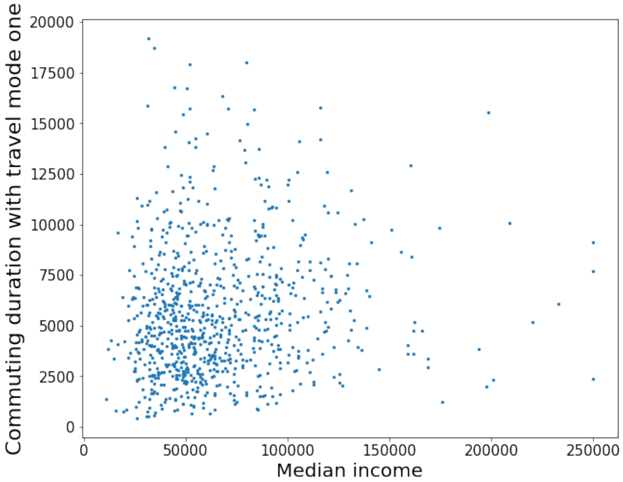

Figure 3: Travel time via mode one v.s. median income: 1,000 OD pairs from the New York–Newark Urban Area.

2. **Time Saving from First/Last Mile Ridesourcing v.s. Income Level (Figure 4).** Given our dataset, the time saving by adopting ridesourcing for first and/or last mile can be obtained by checking the difference between the different travel modes. For example, the relationship between income level and time saving of using ridesourcing to fulfill both first and last mile needs is shown in Figure 4. Each dot in the figure represents one OD pair. This analysis can be enriched in the future by adding ridesourcing fare data, in order to investigate whether the cost of ridesourcing exceeds the benefit of time saving. The figure as shown suggests that the distribution of time saving does not vary much across income levels. This implies a potential transportation inequality, since rich neighborhoods may have more disposable income to utilize ridesourcing services.

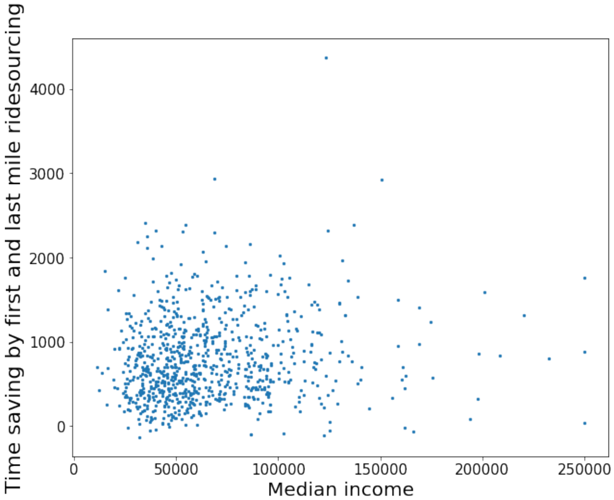

Figure 4: Time saving by adopting ridesourcing for both first and last mile segments v.s. median income: 1,000 OD pairs of New York–Newark Urban Area.

References

[Google Maps Platform, 2021] Google Maps Platform (2021). Google Maps API. `https://developers.google.com/maps`. [Accessed 05-25-2021].

[Reck and Axhausen, 2020] Reck, D. J. and Axhausen, K. W. (2020). Subsidized ridesourcing for the first/last mile: how valuable for whom? *European Journal of Transport and Infrastructure Research*, 20(4):59–77.

[U.S. Census Bureau, 2018a] U.S. Census Bureau (2018a). United States Census Bureau: Cartographic Boundary Files - Shapefile. `https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html`. [Accessed 06-01-2021].

[U.S. Census Bureau, 2018b] U.S. Census Bureau (2018b). United States Census Bureau: Longitudinal Employer-Household Dynamics Datasets. `https://lehd.ces.census.gov/data/`. [Accessed 06-01-2021].

[U.S. Census Bureau, 2019] U.S. Census Bureau (2019). American Community Survey (ACS) Data . `https://www.census.gov/programs-surveys/acs/data.html`. [Accessed 06-12-2021].

[Wang, 2019] Wang, H. (2019). Routing and scheduling for a last-mile transportation system. *Transportation Science*, 53(1):131–147.

[Wang and Odoni, 2016] Wang, H. and Odoni, A. (2016). Approximating the performance of a "last mile" transportation system. *Transportation Science*, 50(2):659–675.