Understanding Travel Demand through Passively-generated Mobile Data: a Python-based Mobility Analysis Workshop

NYCDOT DEV Working Group

Ekin Uğurel, PhD

Department of Technology Management and Innovation (TMI)
NYU Tandon

February 6, 2026

What you will learn today

- Theoretical
  - Basics of travel behavior theory
  - Sources of and issues with large-scale mobility data (and the collection thereof)
  - Introduction to U.S. Census data
  - Some mobility models (i.e., assumptions and limitations)
- Practical
  - Data structures for mobility data
  - Merging with census data
  - Measuring and visualizing mobility quantities (volumes, distances, etc.)
  - Predicting mobility flows with the Gravity model

Note: Some familiarity with coding / object-oriented programming may be beneficial but is not necessary to follow this session.

Overview

Before I start...

Go to this link: https://tinyurl.com/c2smartGit

And hit this button:

What is travel behavior?

(Goulias et al., 2020)

"In this sense, travel behavior is the combination of doing things in different places at different times and how we move from one place to another. Travel behavior is also about feelings, emotions, perceptions, norms, beliefs, intentions, and attitudes. ... Moreover, travel behavior is how to go about deciding how to do things. Perhaps we form utilities for everything we do, or perhaps we use intuition, or perhaps we do both."

"[We] allocate time and other resources to activities and interactions with other people that evolve over time and space."

Why do we care about travel behavior?

Can help planners/engineers answer questions like:

- Where should we place a new transit station?
- How should we tune our signal timing to optimize traffic flow?
- How do we ensure equitable access to employment opportunities given the distribution of work/housing imbalances?
- To what extent can we predict human migration (i.e., moving to a different state for work, international migration, etc.)?

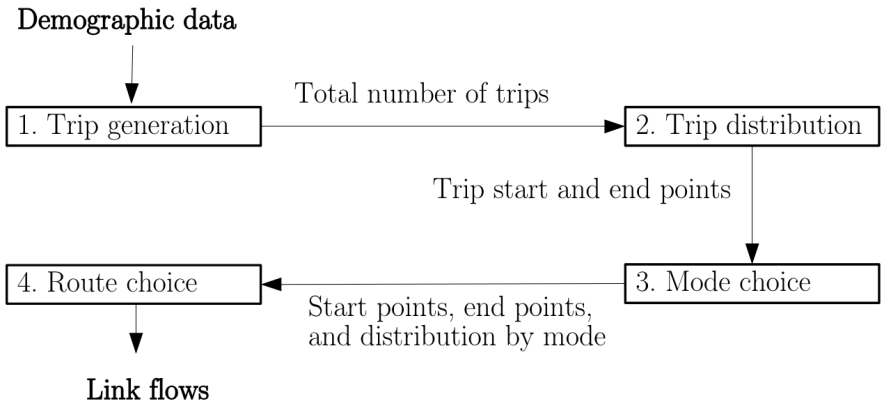
Dimensions of Travel Behavior

- **Who**: The trip maker
- **What**: Trip generation
- **When**: Departure choice, arrival time
- **Where**: Trip distribution, traffic assignment
- **Why**: Trip purpose
- **How**: Mode choice

Influences on Travel Decisions

- Person and household-related attributes
  - Socioeconomics and demographics (McGuckin and Murakami, 1999; Nishii et al., 1988)
  - Attitudes and feelings (Bayarma et al., 2007)
- Built environment
  - Surrounding origin and destination
  - Density, diversity, and design (Cervero and Kockelman, 1997)

The Four-Step Model



Source: (Boyles et al., 2025)

How do we study travel behavior?

We use mobility data, some sources of which include:

- Household travel surveys

How do we study travel behavior?

We use mobility data, some sources of which include:

- Household travel surveys
- Traffic flow counts (i.e., from loop detectors)

How do we study travel behavior?

We use mobility data, some sources of which include:

- Household travel surveys
- Traffic flow counts (i.e., from loop detectors)
- Public transit sensors

How do we study travel behavior?

We use mobility data, some sources of which include:

- Household travel surveys
- Traffic flow counts (i.e., from loop detectors)
- Public transit sensors
- Call detail records

How do we study travel behavior?

We use mobility data, some sources of which include:

- Household travel surveys
- Traffic flow counts (i.e., from loop detectors)
- Public transit sensors
- Call detail records
- **Traces from GPS-equipped devices** ← focus for today

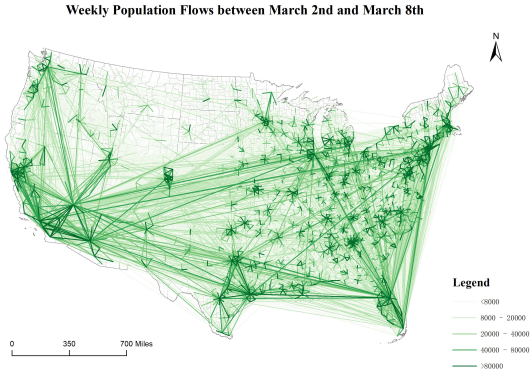
How GPS traces are collected

- Users opt-in to the privacy policies of smartphone apps
- Those apps partner with location data aggregators
- As apps 'ping' the opted-in user's device, GPS data is generated
- Some providers also have access to commercial GPS data (i.e., equipped on long-distance trucks and other commercial vehicles) which tend to be more reliable

Today's Data

Aggregate travel volumes between census block groups (CBGs) for the week of 01/04/2021.

- Publically available here: https://github.com/GeoDS/COVID19USFlows
- Aggregated by SafeGraph

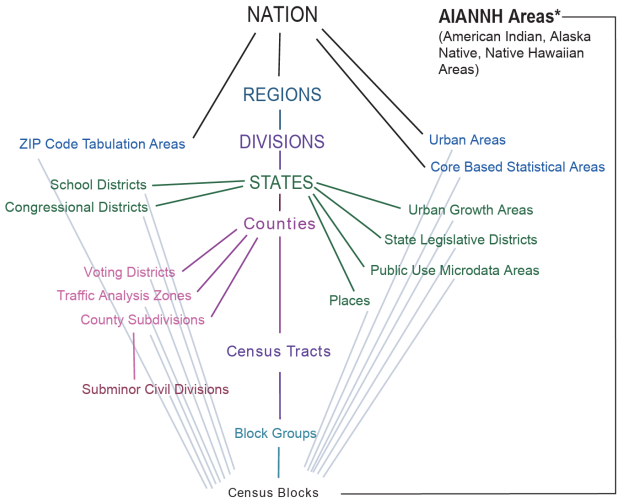


Let's get set up on our code!

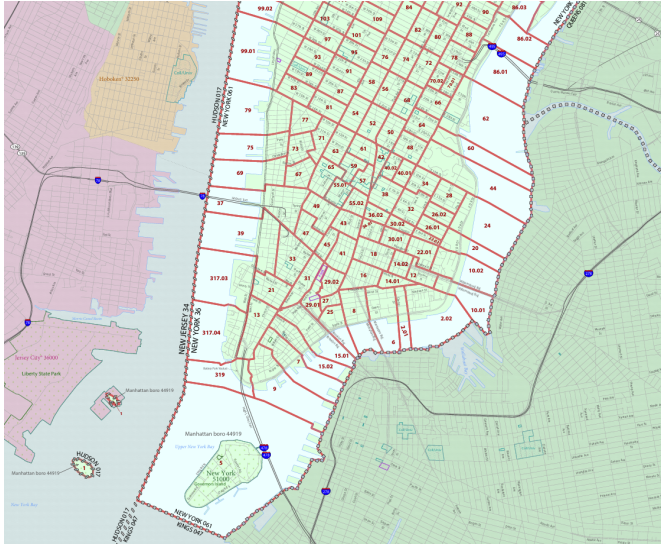
American Community Survey (ACS)

- Conducted every month, every year
- Sent to a sample of addresses (about 3.5 million) in the 50 states, District of Columbia, and Puerto Rico
- Asks about topics not on the 2020 Census, such as education, employment, internet access, and transportation
- Methodology is detailed here.

Census Geography Hierarchy



NYC Census Block Groups



Let's merge ACS data with aggregated mobility data!

Taxonomy of Mobility Models

Individual-level

Population-level

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)

Population-level

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)

Population-level

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)

Population-level

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)
- Other activity-based models (e.g., check out SoundCast!)

Population-level

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)
- Other activity-based models (e.g., check out SoundCast!)

Population-level

- Gravity Model (Zipf, 1946)

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)
- Other activity-based models (e.g., check out SoundCast!)

Population-level

- Gravity Model (Zipf, 1946)
- DNN-based Models (Wu et al., 2024; Rong et al., 2024)

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)
- Other activity-based models (e.g., check out SoundCast!)

Population-level

- Gravity Model (Zipf, 1946)
- DNN-based Models (Wu et al., 2024; Rong et al., 2024)
- Tensor decomposition-based (Li et al., 2023)

Taxonomy of Mobility Models

Individual-level

- Preferential Return (Song et al., 2010)
- Recency (Barbosa et al., 2015)
- Social-based models (De Domenico et al., 2013)
- Other activity-based models (e.g., check out SoundCast!)

Population-level

- Gravity Model (Zipf, 1946)
- DNN-based Models (Wu et al., 2024; Rong et al., 2024)
- Tensor decomposition-based (Li et al., 2023)
- Other data-driven models (Ma et al., 2020)

The Gravity Model

Draws inspiration from Newton's law of gravitational attraction, positing that the flow between two locations is:

- Proportional to the "masses" of the origin and destination (typically population size)
- Inversely proportional to the distance between them

It has the general form:

$$T_{ij} = K \frac{m_i^{\alpha} m_j^{\beta}}{d_{ij}^{\gamma}} \tag{1}$$

where T_{ij} is the flow from origin i to destination j, m_i and m_j are the populations of the origin and destination, d_{ij} is the distance between the origin and destination, and K, α, β, and γ are parameters to be estimated.

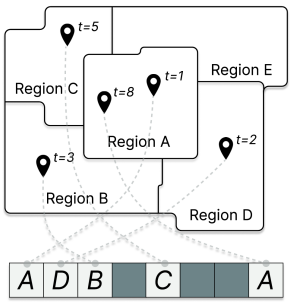
Gravity Model - Parameter Estimation

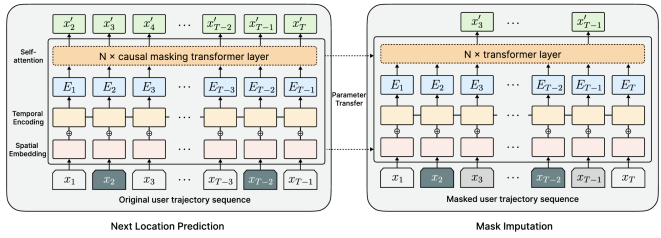
The parameters α, β, and γ are typically estimated from observed flow data. One common approach is to transform the equation into its logarithmic form:

$$\log(T_{ij}) = \log(K) + \alpha \log(m_i) + \beta \log(m_j) - \gamma \log(d_{ij}) \tag{2}$$

This enables the use of Maximum Likelihood Estimation (MLE) to find optimal parameter values that best explain observed flows.

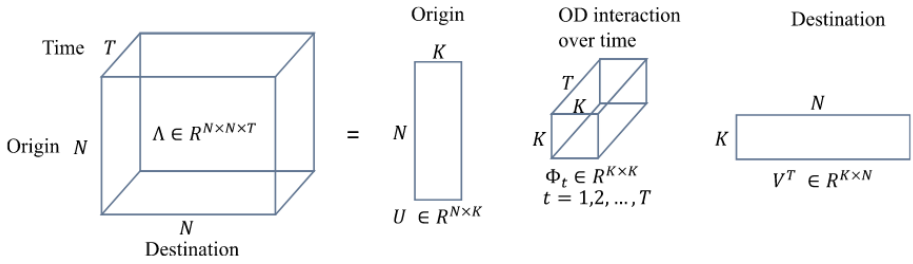
DNN-based Models





Source: (Wu et al., 2024)

Tensor Decomposition-based



Source: (Li et al., 2023)

Model Evaluation

Common Part of Commuters (CPC)

Measures the overlap between predicted and observed flows

$$CPC = \frac{\sum_{i,j} \min(T_{ij}, \hat{T}_{ij})}{\sum_{i,j} T_{ij}} \tag{3}$$

where $\hat{T}_{ij}$ is the predicted flow from zone i to zone j.

- CPC ranges from 0 to 1
- 1 means perfect prediction (all flows match)
- 0 means no overlap between predicted and observed
- Represents the fraction of correctly predicted trips

Model Evaluation

Root Mean Square Error (RMSE)

Quantifies the absolute difference between predicted and observed flows

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,j}(T_{ij} - \hat{T}_{ij})^2} \tag{4}$$

where n is the total number of origin-destination pairs.

- Lower RMSE == better performance
- Sensitive to large errors due to the squared term
- Same unit as flow data, making interpretation straightforward
- Emphasizes absolute errors, which might overemphasize high-flow connections
- For mobility data, often calculated on log-transformed flows to reduce the impact of extreme values

More Resources

- Additional details on LBS data: BigData4Mobility.github.io
- Data Science for Mobility (DSM) Summer School organized by Luca Pappalardo → notebooks here.
- UW Geospatial Data Analysis Course (hosted entirely open-source!)
  - For a more comprehensive treatment of GeoPandas, check out Modules 3, 4, and 6.
- Excellent review of mobility models: *Human mobility: Models and applications*

Our Work

- Uğurel, E., Guan, X., Wang, Y., Huang, S., Wang, Q., and Chen, C. Correcting missingness in passively-generated mobile data with Multi-Task Gaussian Processes. *Transportation Research Part C: Emerging Technologies 161* (Apr. 2024)
- Uğurel, E., Huang, S., and Chen, C. Learning to generate synthetic human mobility data: A physics-regularized Gaussian process approach based on multiple kernel learning. *Transportation Research Part B: Methodological 189* (Nov. 2024), 103064
- Uğurel, E., Wu, X., Wang, R., Lee, B. H. Y., and Chen, C. Metropolitan Planning Organizations' Uses of and Needs for Big Data. *Findings* (Dec. 2024). Publisher: Findings Press
- Wang, Y., Guan, X., Uğurel, E., Chen, C., Huang, S., and Wang, Q. R. Exploring biases in travel behavior patterns in big passively generated mobile data from 11 U.S. cities. *Journal of Transport Geography 123* (Feb. 2025), 104108
- He, J., Sheera, A., Khullar, M., Chavan, S., Herman, B., Uğurel, E., and Mashhadi, A. A framework for measuring and benchmarking fairness of generative crowd-flow models. *ACM Journal on Computing and Sustainable Societies* (To appear in latest edition)

Connect with me!

- I'm interested in solving long-range transportation planning problems using large-scale machine learning (ML).
- My personal website is here: https://ekinugurel.github.io/
- LinkedIn: linkedin.com/in/ekin-ugurel

The End

Please take this anonymous feedback survey to help me make this presentation better

https://tinyurl.com/c2smart

References

Ban, X. J., Chen, C., Wang, F., Wang, J., Zhang, Y., et al. (2018). Promises of data from emerging technologies for transportation applications: Puget sound region case study. Technical report, United States. Federal Highway Administration.

Barbosa, H., de Lima-Neto, F. B., Evsukoff, A., and Menezes, R. (2015). The effect of recency to human mobility. *EPJ Data Science*, 4:1–14.

Bayarma, A., Kitamura, R., and Susilo, Y. O. (2007). Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. *Transportation Research Record*, 2021(1):55–63.

Boyles, S. D., Lownes, N. E., and Unnikrishnan, A. (2025). Transportation network analysis, volume i: Static and dynamic traffic assignment. *arXiv preprint arXiv:2502.05182*.

Cervero, R. and Kockelman, K. (1997). Travel demand and the 3ds: Density, diversity, and design. *Transportation research part D: Transport and environment*, 2(3):199–219.

De Domenico, M., Lima, A., and Musolesi, M. (2013). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1376.

Goulias, K. G., McBride, E. C., and Su, R. (2020). Life cycle stages, daily contacts, and activity-travel time allocation for the benefit of self and others. *Mobility and Travel Behaviour Across the Life Course*, pages 206–220.

Hägerstrand, T. (1970). What about people in regional science. *regional science association*, 24.

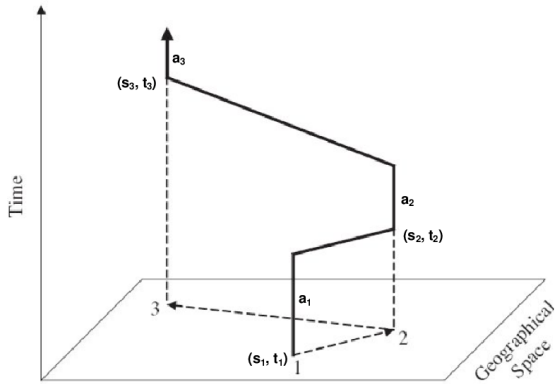
Li, X., Sun, R., Sharpnack, J., and Fan, Y. (2023). Understanding origin-destination ride demand with

Space-Time Geography

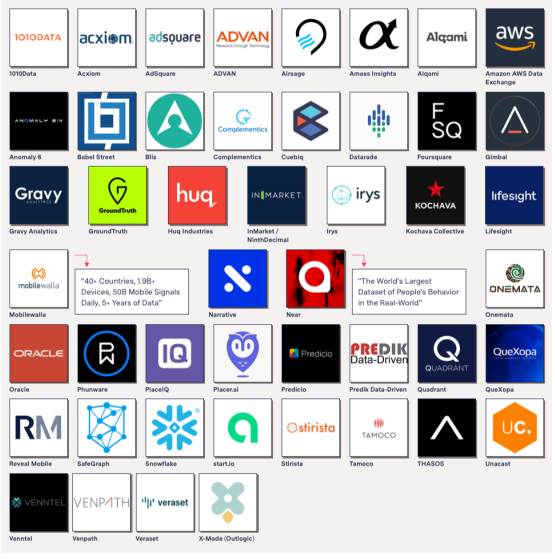
"What about people in regional science?" (Hägerstrand, 1970)

Physical, temporal constraints to locations a person can go.

- Spatial: Origin/destination of trip, travel distance, path chosen, dispersion of trips
- Temporal: Departure time of trips, length of trip, frequency of trips



The Mobility Data Landscape



Source: The Markup

Issues in GPS-based data collection / use

- Sparsity

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy
  - Sensitive information (e.g., one's home and work locations) is easily inferred from high granularity GPS data (De Montjoye et al., 2013)

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy
  - Sensitive information (e.g., one's home and work locations) is easily inferred from high granularity GPS data (De Montjoye et al., 2013)
  - Anonymization and aggregation methods are needed to protect user privacy

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy
  - Sensitive information (e.g., one's home and work locations) is easily inferred from high granularity GPS data (De Montjoye et al., 2013)
  - Anonymization and aggregation methods are needed to protect user privacy
- Bias

Issues in GPS-based data collection / use

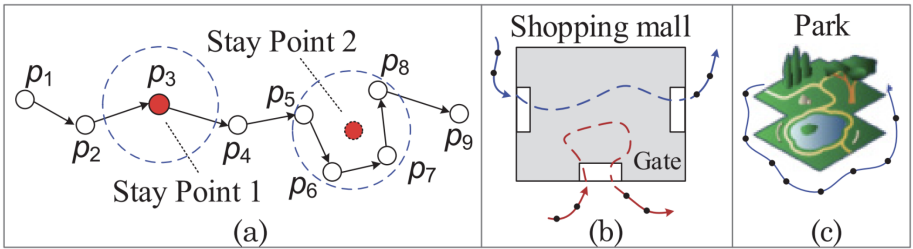
- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy
  - Sensitive information (e.g., one's home and work locations) is easily inferred from high granularity GPS data (De Montjoye et al., 2013)
  - Anonymization and aggregation methods are needed to protect user privacy
- Bias
  - Self-selection bias prevents representativeness (Li et al., 2024)

Issues in GPS-based data collection / use

- Sparsity
  - Peaks and valleys of observation frequency (Ban et al., 2018)
  - 'Urban canyons' & enclosed structures
  - 'Cold start problem'
- Privacy
  - Sensitive information (e.g., one's home and work locations) is easily inferred from high granularity GPS data (De Montjoye et al., 2013)
  - Anonymization and aggregation methods are needed to protect user privacy
- Bias
  - Self-selection bias prevents representativeness (Li et al., 2024)
  - Observed data may distort real-world patterns

Stay Point Detection

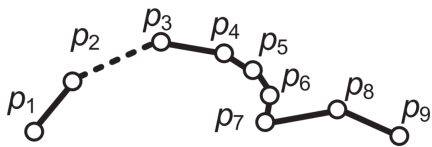
Activity locations where people stay for a period of time.



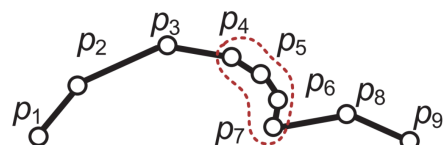
Source: (Zheng, 2015)

Trajectory Segmentation

A trajectory is split into two or more sub- trajectories, with several techniques:



time-interval



stay-point

Source: (Zheng, 2015)