MoveOD: A Data Pipeline for Synthesizing Origin-Destination Demand Data Using U.S. Census Information

VANDERBILT UNIVERSITY

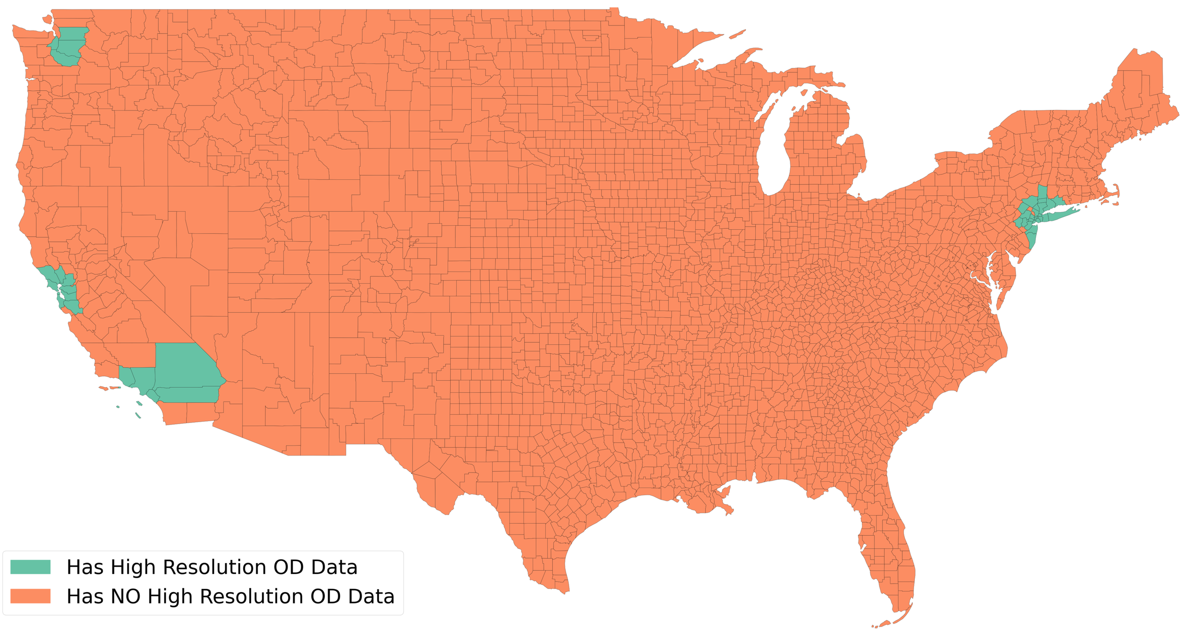
Rishav Sen, Samitha Samaranayake, Ayan Mukhopadhyay, Abhishek Dubey, Aron Laszka

Paper link

Motivation

High-resolution Origin–Destination (OD) data is critical for transportation planning, traffic forecasting, and transit optimization. Yet in the United States, such detailed data remains scarce, except a few cities.

- Accurate OD data critical for traffic modeling, routing, and congestion management.
- Traditional methods (surveys, GPS traces) are insufficient due to sparseness, cost, and bias.
- Roadside sensor data is localized and expensive.



Problem description

The goal of MoveOD is to match public census marginals at multiple levels:

- Total commuters per origin region
- Destination shares from LODES data
- Departure-time block distributions from Census Table B08302
- Travel-time distributions from Census Table B08303

Main Assumption - Independence of distributions: Initial OD assignments assume independence between origin–destination and origin–departure time distributions. This simplifying assumption is explicitly used during the final calibration step to ensure consistency with observed joint patterns.

Features

- **Nationwide Scalability:** Applicable to any U.S. county (using Census data) and other countries with available Census data.
- **Temporal-Spatial Granularity:** Minute-level departure times, building-level spatial resolution.
- **Open-Source & Transparent:** Fully reproducible and adaptable pipeline.
- **Statistical Realism:** Synthesized OD data matches multiple Census data marginals (Census, LODES).

Acknowledgement: This work was supported by the U.S. Department of Energy under Award DE-EE0011188 and the National Science Foundation under Award CNS-1952011.

Data Sources & Pipeline for Point-to-Point OD Generation

Input Data

United States Census Bureau
- Geographic data (region boundaries)
- Aggregated Movement data
  - LODES (for commute)
  - Departure time
  - Travel time

microsoft/GlobalMLBuildingFootpr...
- Generic Building locations
  - Microsoft Buildings footprints

OpenStreetMap
- Residential and Commercial buildings
  - OpenStreetMap (OSM)
- Road speed
  - Default speeds (OSM)

INRIX
  - INRIX (real-time, closed source)

Processing

Adjust LODES to match departure time counts
↓
Spatial sampling (building assignment)
↓
Temporal sampling (minute-level departures)
↓
Travel-time assignment (road-speed based)
↓
Initial OD assignment

Calibration

Integer Linear Program to minimize deviation from Census data. Solved **per origin region** (in our case, census block group):

Objective: $$\min \alpha \sum_{j} (\epsilon_j^+ + \epsilon_j^-) + \beta \sum_{d,s} (\zeta_{o,d,s}^+ + \zeta_{o,d,s}^-)$$

Matches travel-time distribution to Census data — **Penalizes** deviation from initial assignment

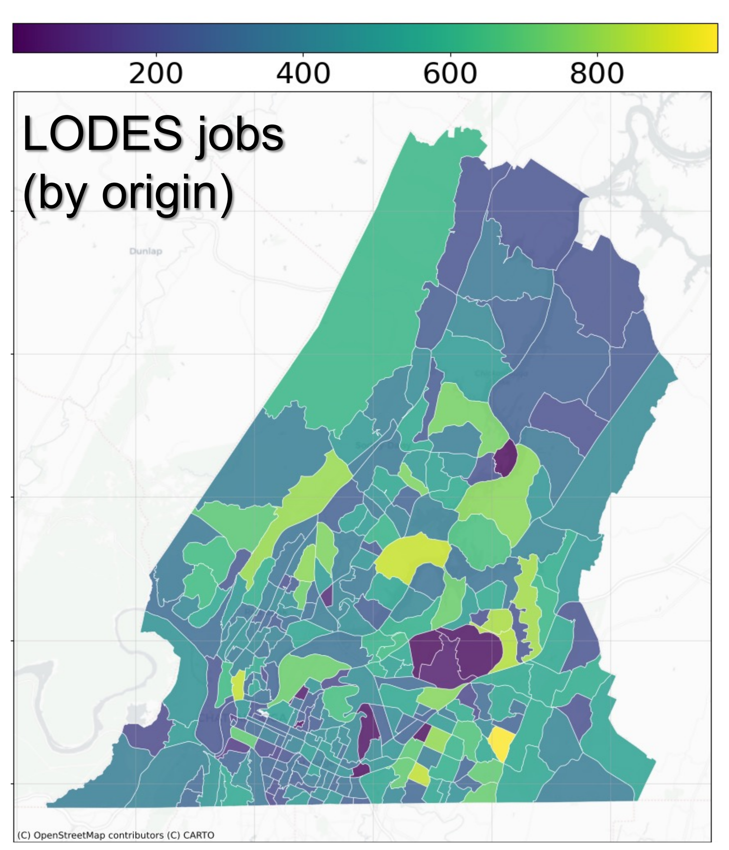
Subject to:

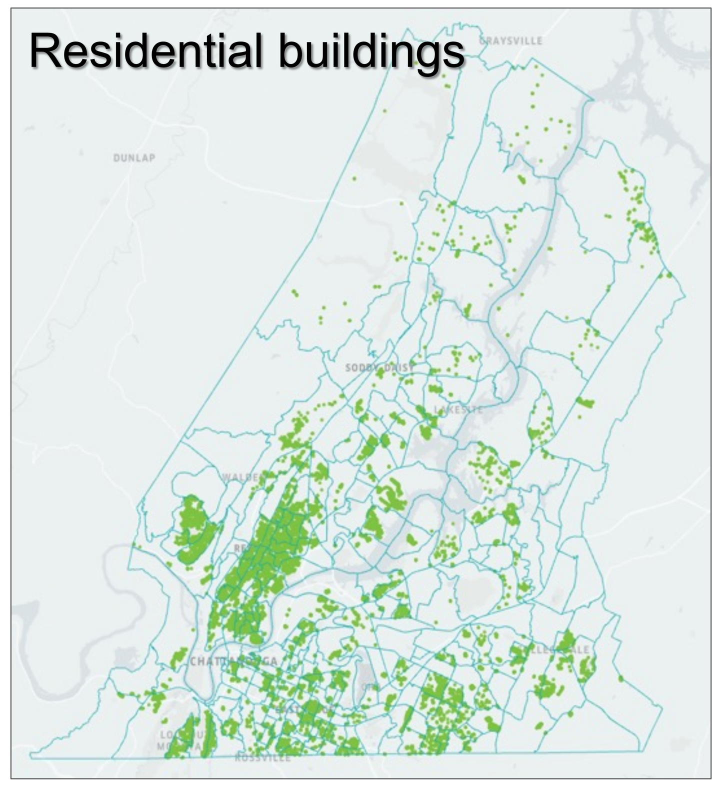
1) $\sum_{(d,t)\in\mathcal{G}\times\mathcal{T}} a_{o,d,t} = N_o$ — Match Total commuters
2) $\forall d \in \mathcal{G}: \sum_{t\in\mathcal{T}} a_{o,d,t} = N_o \cdot \tilde{\lambda}_{o,d}$ — Match destination shares
3) $\forall t \in \mathcal{T}: \sum_{d\in\mathcal{G}} a_{o,d,t} = N_o \cdot \lambda_{o,t}$ — Match departure time shares
4) $\forall j \in \mathcal{J}: \sum_{(d,t)\in\mathcal{G}\times\mathcal{T}} a_{o,d,t} + \epsilon_j^- - \epsilon_j^+ = N_o \cdot \lambda_{o,j}$ — Stay close to Census Travel-times
5) $\forall (d,t) \in \mathcal{G}\times\mathcal{T}: a_{o,d,t} + \zeta_{o,d,t}^- - \zeta_{o,d,t}^+ = m_{o,d,t}$ — Stay close to initial OD

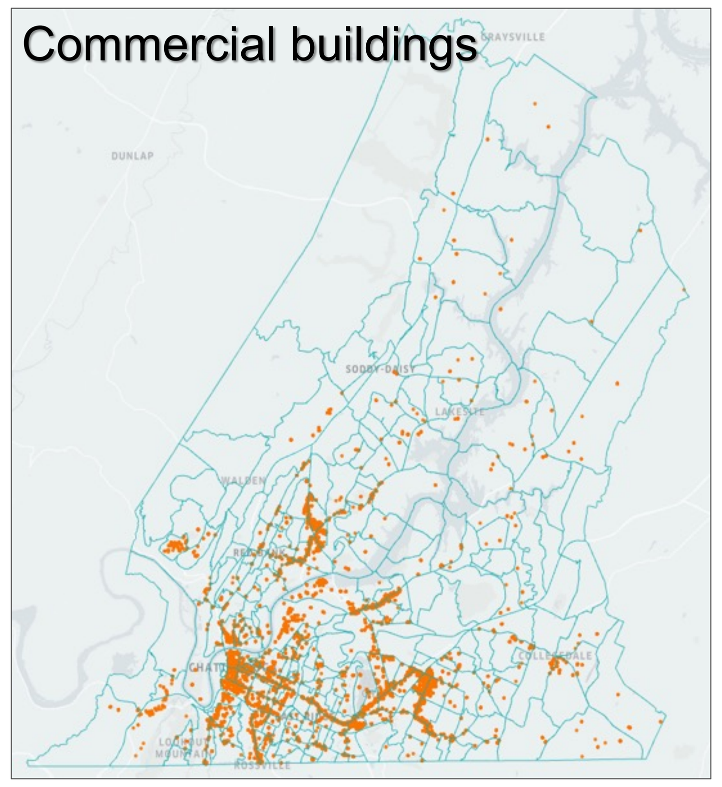
Output

OD pairs with spatial & temporal fidelity

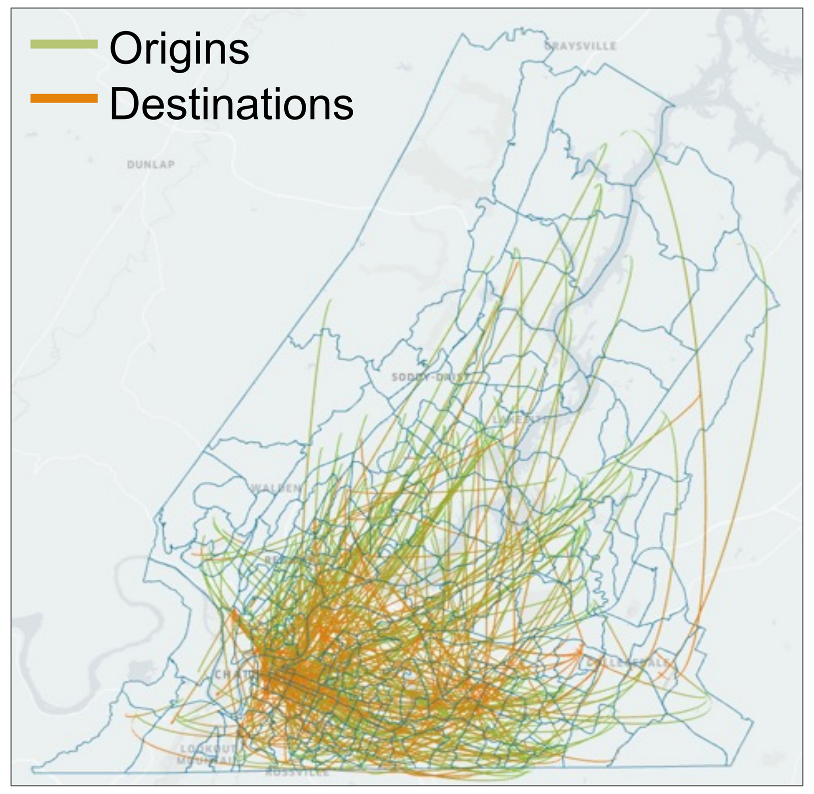
Case Study – Hamilton County, TN



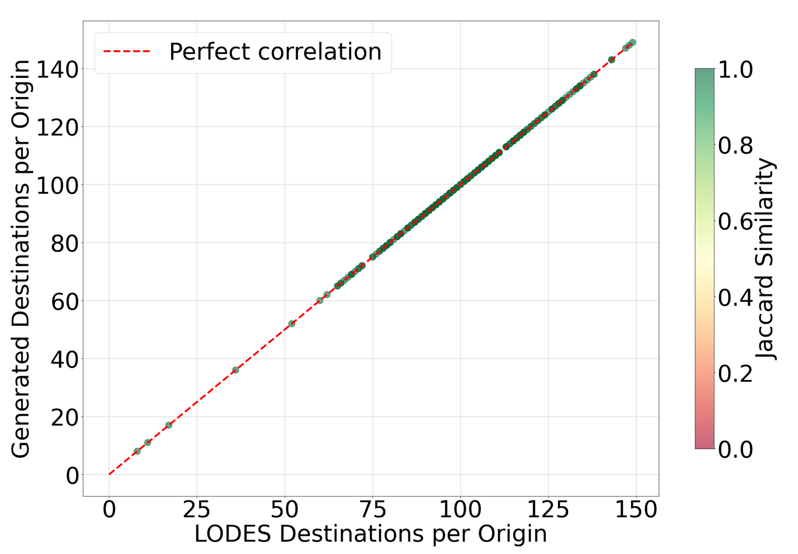




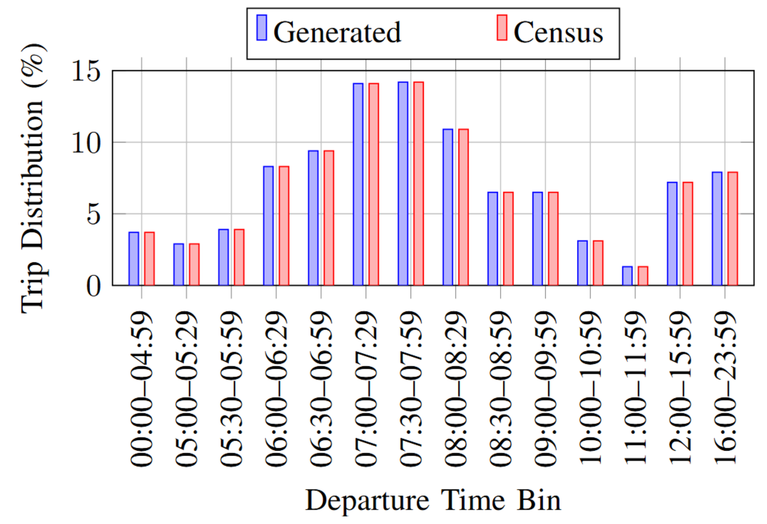
Calibration with commuter counts, departure time and travel time



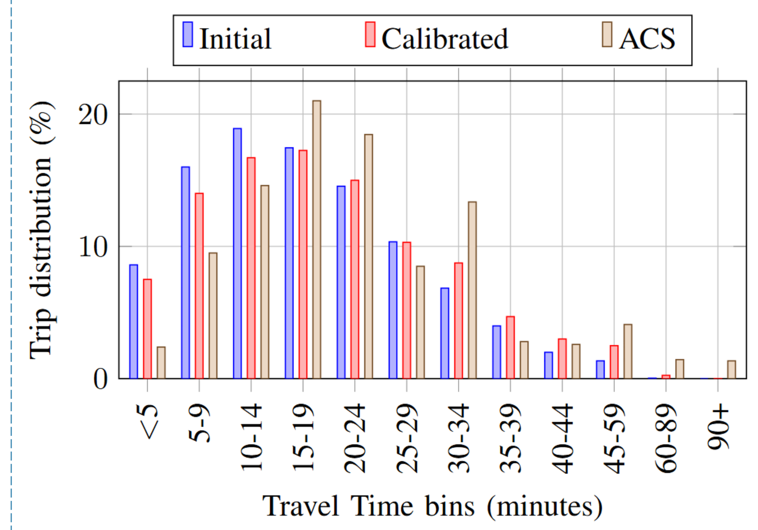
Results & Validation



Jaccard score of 1 indicates perfect preservation of destinations per origin



Census and generated departure times overlap, confirming exact match



Calibrated commute times closely match ACS, preserving shape

[1] Helsgaun, K. (2017). "An Extension of the Lin-Kernighan-Helsgaun TSP Solver for Solving the Generalized Traveling Salesman Problem." *European Journal of Operational Research*
[2] Clarke, G., & Wright, J. W. (1964). "Scheduling of Vehicles from a Central Depot to a Number of Delivery Points." *Operations Research*
[3] Kwon, Y., Kim, J., Kim, J., & Bengio, Y. (2020). "POMO: Policy Optimization with Multiple Optima for Reinforcement Learning." *NeurIPS*

Benchmarking

Using Hamilton County OD data using **30 vehicles** (capacity 4) and **100 trips** to perform capacitated routing while minimizing vehicle miles travelled.

Algorithm	VMT ↓	VMT/PMT ↓	Empty ↓ (%)	Description
LKH-3[1]	2,255	1.55	41	SOTA exact/heuristic hybrid.
Clarke–Wright[2]	2,623	1.80	50	Classical heuristic.
POMO[3]	3,297	2.26	56	Deep RL with policy optimization.

VMT = Vehicle Miles Travelled; VMT/PMT = efficiency ratio; Empty = % distance empty; Coverage = % passengers served.