MoveOD: Synthesizing Origin-Destination Commute Distribution from U.S. Census Data

Rishav Sen*, Jose Paolo Talusan*, Abhishek Dubey*,
Ayan Mukhopadhyay†, Samitha Samaranayake‡, Aron Laszka§
*Vanderbilt University, Nashville, TN, USA
†William & Mary, Williamsburg, VA, USA
‡Cornell University, Ithaca, NY, USA
§Pennsylvania State University, State College, PA, USA

***Abstract*—**
High-resolution origin–destination (OD) tables are essential for a wide spectrum of intelligent transportation systems applications, from modeling traffic and signal timing optimization to congestion pricing and vehicle routing. However, outside a handful of data rich cities, such data is rarely available. We introduce MOVEOD, an open-source pipeline that synthesizes public data into commuter OD flows with fine-grained spatial and temporal departure times for any county in the United States. MOVEOD combines five open data sources – American Community Survey (ACS) departure time and travel time distributions, Longitudinal Employer–Household Dynamics (LODES) residence-to-workplace flows, county geometries, road network information from OpenStreetMap (OSM), and building footprints from OSM and Microsoft, into a single OD dataset. We use a constrained sampling and integer-programming method to reconcile the OD dataset with data from ACS and LODES. Our approach involves: (1) matching commuter totals per origin zone, (2) aligning workplace destinations with employment distributions, and (3) and calibrating travel durations to ACS-reported commute times. This ensures the OD data accurately reflects commuting patterns. We demonstrate the framework on Hamilton County, Tennessee, where we generate roughly 150,000 synthetic trips in minutes, which we feed into a digital-twin simulator with a benchmark suite of classical and learning-based vehicle-routing algorithms. The MOVEOD pipeline is an end-to-end automated system, enabling users to easily apply it across the United States by giving only a county and a year; and it can be adapted to other countries with comparable census datasets. The source code to a lightweight browser dashboard will be made publicly available.

***Index Terms*—Origin-Destination Synthesis, Synthetic Data Generation, Transportation Systems, Data Fusion, Urban Mobility**

I. INTRODUCTION

Granular origin-destination (OD) travel demand data is foundational for intelligent transportation networks and cyber-physical systems (CPS), enabling dynamic traffic management, predictive routing, and adaptive signal control systems [1], [2], [3]. A central challenge in large-scale societal CPS is that feedback, control, and learning pipelines are increasingly data-driven, yet high-fidelity, representative real-world data is often unavailable [4], [5]. Without community-representative demand data, transportation CPS research frequently relies on datasets such as NYC taxi traces that do not reflect the target community, limiting the generalizability of resulting models and control policies. However, collecting fine-grained, building-level OD data remains a persistent challenge. Traditional methods such as household travel surveys, Bluetooth MAC trackers, or mobile GPS traces are often sparse, noisy, or available only for specific travel modes. These sources struggle to generalize across urban, suburban, and rural environments, posing a significant challenge for the over 3,000 smaller transit agencies across the United States that lack integrated data pipelines [6].

Several publicly available nationwide datasets capture aspects of commuting, but none provides a complete distribution. The American Community Survey (ACS) [7] and Longitudinal Employer-Household Dynamics (LODES) [8] provide aggregate counts of residents and jobs by census unit, but lack departure times and fine-grained (i.e., building-level) locations. Building footprints from OpenStreetMap (OSM) [9] or Microsoft (MSBF) [10] pinpoint every structure but contain no occupancy or travel-demand information. Traffic datasets (e.g., INRIX) provide fine-grained demand aggregated over road segments, but cannot be decomposed into individual OD trips. Naively combining these marginal datasets (e.g., via iterative proportional fitting) may yield implausible results, such as hundreds of commuters assigned to a small residential building or every commuter departing in a single rush-hour spike.

These limitations motivate a synthetic approach. In this work, we introduce MOVEOD, an open-source framework for synthesizing granular, time-dependent OD datasets by integrating heterogeneous data sources, including census marginals, employment flows, and road networks, with physical constraints such as building locations, spatial coherence, and time-dependent travel feasibility [11], [12]. While MOVEOD produces synthetic data, it maintains statistical fidelity by matching observed marginals (residence-workplace locations, departure times, travel durations) while enforcing spatial coherence. The result is a CPS data and calibration layer that supports realistic closed-loop evaluation in simulation, optimization, and learning-based control, enabling downstream works to validate their pipelines against community-representative demand rather than geographically mismatched proxies. We focus specifically on daily commuting trips between residences and workplaces, allowing MoveOD to leverage spatial employment and housing density data while avoiding poorly measured recreational or freight travel.

Our core contribution is not a static dataset but a modular synthesis approach: MoveOD can generate OD datasets for any U.S. county and, in principle, for any region where similar marginal datasets are available. We release an interactive platform to generate commute OD data and demonstrate its utility on Hamilton County and Davidson County in Tennessee, US.

II. Related Works

Open Benchmarks for Transportation AI: The past few years have witnessed remarkable growth in large-scale, openly licensed datasets that have catalyzed progress in traffic forecasting, routing, and autonomous driving research. Early sensor-centric research such as METR-LA and PEMS-BAY [13] sparked the first wave of spatio-temporal graph neural networks; subsequent releases expanded in size (e.g., LargeST [14]) or semantic richness (e.g., ScenarioNet [15]). While indispensable, these datasets share two limitations that are critical for OD research: *(i)* they focus on traffic states measured at fixed sensor locations, not on the population-level flows between origins and destinations; *(ii)* coverage is typically confined to a few metropolitan areas, hindering work on domain transfer and equitable deployment across diverse cities.

Origin–Destination Estimation:

Several open projects attempt full synthetic generation. Open-PFLOW [16] uses stochastic methods to allocate commuting trips to census microdata, providing disaggregate, time-resolved movement data that enables analysis of flows at different times of day. However, its extensibility beyond Japan is limited, and the generation process can rely on private or commercial datasets for detailed spatial information, such as building locations and road networks. The vehicle-network simulator of [17] produces microscopic trajectories for Vehicular Ad Hoc Network (VANET) studies, yet its demand model is tuned to a single European city and lacks transparency in demographic assumptions.

Simulators for Transportation Systems: Comprehensive microsimulation platforms like SUMO [18] and VISSIM [19] are widely used in transportation research. However, they present significant limitations for deep reinforcement learning (DRL) applications in intelligent transit dispatching problems. These simulators are designed for detailed traffic analysis and model individual vehicle dynamics, signal timing, and microscopic interactions that create substantial computational overhead during DRL training, which typically requires thousands to millions of episodes.

Public-Transit and Multi-Modal Planning: Modern, data-driven transit planning uses OD matrices to size fleets and schedule services [20], [21]. Having finer-grained OD data improves the calibration of discrete-choice models [22] and supports more accurate multi-modal assignment [23]. However, most public transit simulations [24], [25] assume a fixed commuter OD table obtained from proprietary surveys, limiting reproducibility and flexibility.

III. Problem Statement and Data Sources

Our objective is to infer a fine-grained joint distribution of commute data at the level of individual buildings and minutes of the day. In our setting, origins are residences and destinations are workplaces.

Spatial sets Let $\mathcal{G}$ be the set of census units (CUs) and $\mathcal{B}$ be the set of all buildings. For each $g \in \mathcal{G}$, we define its building set

$$B_g = \{\, b \in \mathcal{B} : G(b) = g\},$$

where $G : \mathcal{B} \to \mathcal{G}$ maps a building b to the census unit $G(b)$ that contains the building.

Origin and destination notation Each CU can serve as both an origin and a destination. Let $o \in O \subseteq \mathcal{G}$ denote origin CUs and $d \in D \subseteq \mathcal{G}$ denote destination CUs. Specific buildings in the origin and destination CUs are written as $b_o \in B_o, b_d \in B_d$, where b_o and b_d are sampled from the set of buildings in their respective census units.

Temporal distribution Each day is indexed by minutes $t \in \mathcal{T}$. The ACS departure-time blocks are $s \in \mathcal{S}$ where $s = [t_l^s, t_u^s)$ represents an interval in hours (e.g., $[0, 6)$ for "Before 6 AM" and $[6, 7)$ for "6–7 AM"). To generate minute-level outputs, we sample a departure minute m_o uniformly from the corresponding minute indices in the selected hour block, i.e., $m_o \in \{t_l^s, \dots, t_u^s\}$. Travel times are represented by bins $k \in \mathcal{K}$ (e.g., 0–5 minutes, 5–10 minutes, and so on).

Commuters The total commuters per origin are represented as N_o for $o \in O$. The random vector describing a commuter's trip is $\big(B_o, B_d, M_o, M_d\big)$, where

$$\begin{aligned} B_o &\in \mathcal{B} && \text{is the origin building,}\\ B_d &\in \mathcal{B} && \text{is the destination building,}\\ M_o &\in \mathcal{T} && \text{is the departure time from origin,}\\ M_d &\in \mathcal{T} && \text{is the arrival time at destination.} \end{aligned}$$

A specific commuter is represented as (b_o, b_d, m_o, m_d), and our goal is to estimate $P\big(B_o = b_o, B_d = b_d, M_o = m_o, M_d = m_d\big)$,

Notation conventions. We use lowercase symbols (o, d, m_o, m_d) as indices. The values m_o and m_d are minute-level realizations of the random variables S (departure time) and E (arrival time), respectively. We use uppercase symbols (O, D, S, E) for random variables, $n.$ for counts, and $p.$ for normalized probabilities.

A. Public Data Sources

Our framework relies exclusively on publicly available datasets to provide the spatial and temporal inputs needed for OD synthesis. Table I provides a summary of the data sources used. Below, we outline how we acquire and pre-process each dataset.

Census unit (CU) To obtain the boundary of each CU, we download the U.S. Census Bureau's TIGER/Line shapefiles for block groups (or another desired summary level) directly

TABLE I: Data sources

Symbol	Type	Description
$\langle O, D\rangle$	LODES	Origin-destination flow data
$\langle O, T\rangle$	ACS B08302	Departure time by origin census unit
$\langle O, J\rangle$	ACS B08303	Travel time distribution by origin CU
$\mathcal{B}$	OSM/MSBF	Building footprints (residence/workplace)
$\mathcal{V}$	INRIX/OSM	Road speed data (hourly)

from the Census website. These open-source TIGER/Line files provide precise polygon boundaries for every block group in the United States.

Commuter Origin-Destination The LEHD Origin–Destination Employment Statistics (LODES) provides a joint distribution of commuter counts for each origin–destination CU pair [26]. We represent $n_{o,d}$ as the count of commuters with origin $o \in O$ and destination $d \in D$.

Then the estimated probability mass function for the (o, d) pair, conditioned on the origin, is $p_{o,d} = \frac{n_{o,d}}{\sum_{d' \in D} n_{o,d'}}$.

Departure Time Marginals (ACS Table B08302) Let $\mathcal{C}_{\text{dep}}$ be the ACS commuter records from Table B08302. We define $n_{o,s}$ as the count of commuters originating at $o \in O$ with a departure time-block $s \in \mathcal{S}$. By construction, $\sum_{s \in \mathcal{S}} n_{o,s} = N_o$, i.e., the total weighted commuter count at origin o. We then set $p_{o,s} = \frac{n_{o,s}}{N_o}$, treating $p_{o,s}$ as the estimated origin–departure marginal, i.e., the population share leaving origin o in each departure-time block s.

Travel–Time Marginals (ACS Table B08303) Let $\mathcal{C}_{\text{tt}}$ be the weighted set of travel-time records in ACS B08303. The travel-time bin $k \in \mathcal{K}$ indicates the time taken for a commuter to reach their destination. ACS provides the joint marginal only between origin and travel time. For each origin CU $o \in O$ and travel-time bin k, ACS B08303 reports a count $n_{o,k}$. We obtain the empirical joint distribution $p_{o,k} = \frac{n_{o,k}}{\sum_{k' \in \mathcal{K}} n_{o,k'}}$. which serves as our estimate of the travel-time distribution conditioned on origin.

Tiered building selection. We start by identifying two subsets of $\mathcal{B}$. $\mathcal{B}_{\text{OSM}}$ contains buildings with known locations and OpenStreetMap land-use tags $L(b)$ (residential, commercial, industrial, or other). OSM coverage can be incomplete. The second subset is $\mathcal{B}_{\text{MSBF}}$, from Microsoft's GlobalMLBuilding-Footprints data, which provides broad building coverage but does not include land-use tags. If a census unit has no building data in either source, we use the CU centroid $b_{\text{centroid}}(g)$ as the fallback building location.

For each origin CU $o \in O$, the candidate building set is

$$B_o = \begin{cases} \{b \in \mathcal{B}_{\text{OSM}} : G(b) = o, L(b) = \text{res}\}, & \text{if} \neq \varnothing, \\ \{b \in \mathcal{B}_{\text{MSBF}} : G(b) = o\}, & \text{else if} \neq \varnothing, \\ \{b_{\text{centroid}}(o)\}, & \text{otherwise} \end{cases}$$

For each destination CU $d \in D$, the candidate destination-building set is

$$B_d = \begin{cases} \{b \in \mathcal{B}_{\text{OSM}} : G(b) = d, L(b) = \text{com}\}, & \text{if} \neq \varnothing, \\ \{b \in \mathcal{B}_{\text{MSBF}} : G(b) = d\}, & \text{else if} \neq \varnothing, \\ \{b_{\text{centroid}}(d)\}, & \text{otherwise} \end{cases}$$

Road speeds. All road segments $z \in \mathcal{Z}$ have a speed $v \in \mathcal{V}$. The speed is determined using:

- **OpenStreetMap (OSM) defaults:** Provides the road network and each road segment's speed limit, or a standard average speed based on the type of road.
- **INRIX data (proprietary, optional):** Provides hourly historical road speeds using average measurements from roadside sensors.

We use INRIX speed values for roads where INRIX data is available; for other roads, we use speed values from OSM. This hybrid approach yields a granular travel-time profile $\tau_z(t)$ for road segment $z \in \mathcal{Z}$ at all times $t \in \mathcal{T}$. The availability of real-time road speeds makes synthetic data generation and calibration more accurate.

IV. Approach

Having outlined the dataset requirements, we next detail the statistical integration process used to generate the commute OD representation.

Bayesian Decomposition. Our target is to generate the complete joint distribution $P(O, D, S, E)$, where O is the origin CU, D is the destination CU, S is departure time, and $E = S+K$ is arrival time. By the chain rule, $P(O, D, S, E) = P(E \mid S, D, O)\, P(S \mid D, O)\, P(D \mid O)\, P(O)$. We simplify with the following assumption: Departure-time blocks S depend only on the origin: $P(S \mid D, O) = P(S \mid O)$. Thus $P(O, D, S, E) = P(E \mid S, D, O)\, P(S \mid O)\, P(D \mid O)\, P(O)$. We estimate each of these marginals as follows:

- $P(O)$ from the CUs in a county.
- $P(D \mid O)$ from LODES, and then assigning each commuter an origin and destination building.
- $P(S \mid O)$ from departure-time marginals (ACS B08302).
- $P(E \mid S, D, O)$ by combining road-network travel times with ACS travel-time marginals (Table B08303).

Combining these steps yields a fully specified $P(O, D, S, E)$ that is consistent with available public distributions and can be sampled to produce building- and minute-level synthetic OD trips. Keeping $P(D \mid O)$ and $P(S \mid O)$ intact anchors the synthetic data to observed spatial and temporal commuting patterns. These data sources give us four anchors:

1) Origin-to-destination counts $n_{o,d}$.
2) Building location sets B_o and B_d represent the residential and commercial buildings, respectively, in every CU.
3) Conditional departure-time distribution $p_{o,s}$.
4) Conditional travel-time distribution $p_{o,k}$.

The methodology below turns these anchors into minute-level, building-specific OD assignments $\mathcal{A} = \{(b_o, b_d, m_o, m_d)\}$, which are then calibrated against the travel-time distribution in ACS Table B08303.

1. Spatial Sampling Within each census unit, we assign origin and destination buildings uniformly to each of the $n_{o,d}$ commute trips, prioritizing tagged buildings over untagged ones. For each origin CU, we select an origin building $b_o \in B_o$; for destination CU d, we select $b_d \in B_d$.
2. Departure-Time Assignment Given an origin CU o and departure-time block $s \in \mathcal{S}$ (with minute bounds t_l^s, t_u^s), we sample a minute-level departure time uniformly within the corresponding minute range: $m_o \sim \text{Unif}(\{t_l^s, \ldots, t_u^s\})$. This preserves the origin and departure-time marginals by construction.
3. Arrival at Destination For each sampled tuple (b_o, b_d, m_o), we compute:

- route length $\ell_{o,d} = \text{graph-distance}(b_o, b_d)$,
- shortest-path travel time $\tau_{o,d}(m_o) = \text{duration}(b_o, b_d \mid m_o)$,
- implied speed $v_{o,d}(m_o) = \ell_{o,d}/\tau_{o,d}(m_o)$.

Define

$$\Omega = \{\tau_{o,d}(m_o)\}$$

over sampled trips. The arrival minute is

$$m_d = m_o + \tau_{o,d}(m_o).$$

Hence the initial granular OD dataset is

$$\mathcal{A} = \{(b_o, b_d, m_o, m_d)\}.$$

Initial OD Dataset The dataset $\mathcal{A}$ satisfies ACS origin totals, LODES origin-destination marginals, and ACS departure-time marginals by construction, but still requires calibration to ACS travel-time marginals.
Mean Road Speed Shift To correct for any bias in travel-time reporting in ACS Table B08303, we perform a mean road-speed shift. Self-reported travel times usually overestimate actual travel times because of rounding errors and the inclusion of parking or waiting times [27], [28].

Let $N = \sum_{o \in O} N_o$ be the total number of commuters. The mean travel time from the initial OD dataset is

$$\bar{\tau}^{\text{init}} = \frac{1}{N} \sum_{\tau \in \Omega} \tau.$$

Let $\bar{\tau}^{\text{ACS}}$ be the ACS mean commute time from Table B08303, computed from travel-time bin midpoints. We use the multiplicative speed-shift factor

$$\psi = \frac{\bar{\tau}^{\text{init}}}{\bar{\tau}^{\text{ACS}}},$$

and adjust road-segment speeds by $v'_r = \psi\, v_r$.

V. Calibration

The goal is to calibrate $\mathcal{A}$ to match ACS travel-time marginals while preserving origin-destination and origin-departure marginals.
Inputs for a fixed origin CU $o \in O$

1) $m_{o,d,s} \in \mathbb{Z}_{\geq 0}$: commuters in $\mathcal{A}$ from origin o to destination d in departure block $s \in \mathcal{S}$.
2) $N_o \in \mathbb{Z}_{>0}$: total commuters whose home CU is o.
3) $p_{o,d} \geq 0$, $\sum_{d \in D} p_{o,d} = 1$: destination-CU distribution.
4) $p_{o,s} \geq 0$, $\sum_{s \in \mathcal{S}} p_{o,s} = 1$: departure-time distribution (ACS B08302).
5) $p_{o,k} \geq 0$, $\sum_{k \in \mathcal{K}} p_{o,k} = 1$: travel-time-bin distribution (ACS B08303).
6) $\pi_{o,d,s,k} \in [0,1]$: fraction of commuters in cell (o,d,s) whose travel time falls in bin k.

Decision variables

1) $a_{o,d,s} \in \mathbb{Z}_{\geq 0}$: calibrated commuters assigned to (d,s).
2) $\eta_k^+, \eta_k^- \geq 0$: slacks for travel-time bin $k \in \mathcal{K}$.
3) $\zeta_{o,d,s}^+, \zeta_{o,d,s}^- \geq 0$: slacks from initial counts $m_{o,d,s}$.

A. Integer linear program

Objective:

$$\min\ \alpha \sum_{k \in \mathcal{K}} (\eta_k^+ + \eta_k^-) + \beta \sum_{(d,s) \in D \times \mathcal{S}} (\zeta_{o,d,s}^+ + \zeta_{o,d,s}^-)$$

Subject to:

1) $\displaystyle\sum_{(d,s) \in D \times \mathcal{S}} a_{o,d,s} = N_o$
2) $\forall d \in D:\ \displaystyle\sum_{s \in \mathcal{S}} a_{o,d,s} = N_o\, p_{o,d}$
3) $\forall s \in \mathcal{S}:\ \displaystyle\sum_{d \in D} a_{o,d,s} = N_o\, p_{o,s}$
4) $\forall k \in \mathcal{K}:\ \displaystyle\sum_{(d,s) \in D \times \mathcal{S}} \pi_{o,d,s,k} a_{o,d,s} + \eta_k^- - \eta_k^+ = N_o\, p_{o,k}$
5) $\forall (d,s) \in D \times \mathcal{S}:\ a_{o,d,s} + \zeta_{o,d,s}^- - \zeta_{o,d,s}^+ = m_{o,d,s}$

Constraints (1)–(3) enforce destination and departure marginals exactly. Constraint (4) matches travel-time marginals up to slacks η_k^+, η_k^-, and constraint (5) keeps calibrated counts close to the initial OD dataset.

The calibrated OD counts $\{a_{o,d,s}\}$ are mapped back to buildings by resampling within each (o,d,s) cell and sampling departure minutes inside each block s, then setting arrival time as

$$m_d = m_o + \tilde{\tau}_{o,d}(m_o),$$

where $\tilde{\tau}_{o,d}(m_o)$ is the mean-speed-shifted travel time.

VI. Experimental Setup & Results

We demonstrate that the pipeline behaves as intended by applying it to Davidson and Hamilton counties in Tennessee, U.S. With more than 336,000 commuters, Davidson County is the largest county in Tennessee. We later use the Hamilton County OD dataset in multiple benchmarks. All experiments were performed on a 32-core, 5.2 GHz, 32 GB RAM Unix machine.

For the α, β hyperparameters in the calibration step, we sweep both over $[0,1]$ in increments of 0.1. Experimentally, $\alpha = 1$ and $\beta = 1$ provide the smallest combined gap between the initial OD dataset $\sum_{(d,s)} (\zeta_{o,d,s}^+ + \zeta_{o,d,s}^-)$ and the ACS travel-time distribution $\sum_{k \in \mathcal{K}} (\eta_k^+ + \eta_k^-)$.
Validating the marginals. Figure 1 overlays the ACS departure-time histogram with the synthetic departures; the two curves coincide, confirming that the block-level departure

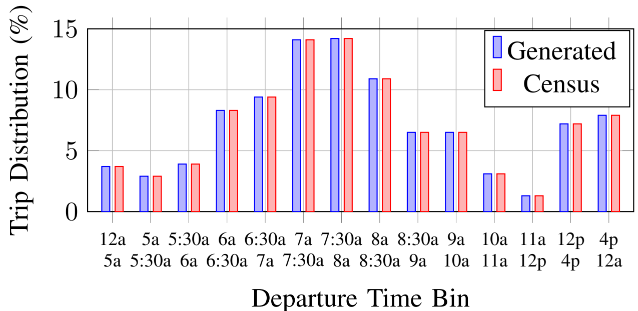


Fig. 1: MoveOD calibrates the departure times for all origin census units to align with ACS departure times (B08302).

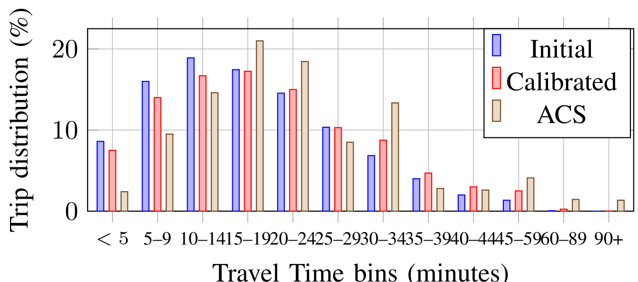


Fig. 2: Travel time distributions: initial commuter assignment (Initial), calibrated commuter assignment (Calibrated), and ACS travel time distribution (Table B08303).

times are reproduced exactly. Figure 2 compares the distribution of commute travel times across the initial commuter assignment (in blue), the calibrated commuter assignment (in red), and the ACS Table B08303 distribution (in brown). The calibrated synthetic data are more closely aligned with the ACS distribution than the initial synthetic data, both in the left tail (up to 15 minutes) and in the right tail (20 to 90 minutes), while preserving the mode and overall shape. However, the calibrated dataset slightly underestimates the proportion of typical travel times (15–19 minutes) compared to ACS. The calibration process thus improves fidelity to observed census patterns, but some discrepancies remain around the mean of the distribution. Overall, the results indicate strong structural preservation and successful calibration for most commuters.

Principally, the Hamilton County experiment shows that MoveOD (i) preserves the origin-destination and origin-departure marginals, (ii) improves alignment with ACS travel-time marginals through calibration, and (iii) exposes a tunable parameter to account for survey bias in travel times.

Time Complexity. The main cost scales linearly with the number of trips M. Spatial sampling and departure-time assignment are $O(M)$; travel-time computation is $O(M \cdot \log Y)$, where Y is the number of nodes. Speed adjustment is $O(M + Z)$, with Z road edges. Overall runtime is $O(M \cdot \log Y + Z)$. MoveOD runs in 22 minutes for ∼336K commuters (Davidson County, TN) and 14 minutes for ∼150K commuters (Hamilton County, TN).

Space Complexity. Memory use is $O(N)$ for commuter assignments, $O(B)$ for building tables, $O(G \cdot (S + K))$ for ACS marginals, and $O(Z + Y)$ for the road network. Here, N is the number of commuters, B is the number of buildings, G is the number of census units, S is the number of departure-time bins, K is the number of travel-time bins, Z is the number of road segments, and Y is the number of network nodes. Total space complexity is $O(N + B + G \cdot (S + K) + Z + Y)$.

A. Interactive Interface and Digital Twin

Data Generation. MoveOD uses a minimalist web interface, Fig. 3a, where a user selects the state, county, range of dates to synthesize, the LODES release year, and an optional INRIX road-speed feed. The number of points has been truncated for visibility. Other inputs, such as ACS tables, OSM, MSBF, and CU boundary files, are retrieved automatically. Pressing `BEGIN` returns ready-to-use calibrated OD tables, building metadata, and shapefiles.

Simulation. The digital twin can ingest different user-demand requests, including those generated by MoveOD, to simulate a range of transit configurations, from fully fixed-line service and on-demand service to multimodal systems that use optimization algorithms to identify the best mode for each user at any time. Figures 3b and 3c show how the twin can load existing General Transit Feed Specification (GTFS) and/or on-demand configurations. The tool allows operators or analysts to switch between systems and adjust configurations to evaluate performance for the user demand shown in Fig. 3d.

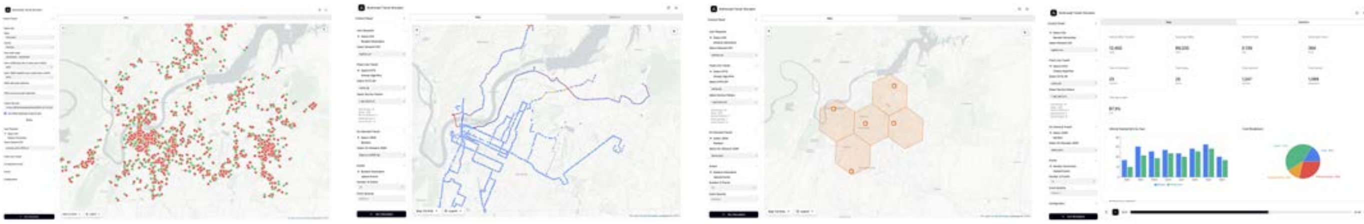
Open Source. A browser-based, front-end tool will be made available at https://github.com/scope-lab-vu/moveOD. This repository lets users deploy a visual dashboard that allows them to choose a state, county, date, and optional spatial or temporal filters, then generate calibrated OD data in minutes.

VII. Discussion and Conclusions

MoveOD delivers building- and minute-level commuter flows by fusing LODES OD distributions, ACS departure-time and travel-time distributions, and building footprints from OpenStreetMap and Microsoft Building Footprints. Unlike prior efforts that consider either where or when people commute, MoveOD matches both spatial and temporal dimensions to ground-truth marginal distributions, while remaining fast enough to apply to any U.S. county.

Typical applications. OD data show where people commute for work, which is one of the primary drivers of peak traffic conditions. OD data can be used for applications such as (i) predicting county-level traffic flows, (ii) public-transit design, and (iii) efficient road-network design. Moreover, the data generation framework is designed to be flexible: it can ingest any macro-level movement dataset and enhance spatial resolution to building-level origins and destinations. If temporal distributions are available, the process further refines the OD dataset with time-aware commuter assignments. This adaptability allows integration with diverse data sources, such as crowd-sourced GPS trajectories or emerging Internet of Things (IoT) mobility streams.

Limitations and outlook. MoveOD currently generates OD data only for weekday commute trips. Future work will add weekend and non-commute trips, account for weather and seasonal effects, and ingest real-time traffic speed data.



(a) Synthetic MoveOD Data: Residence to Workplace

(b) Public Transit routes and stops across Hamilton County

(c) Sample On-Demand Service Zones and Depot

(d) Metrics for multimodal transit given MoveOD requests

Fig. 3: MOVEOD data generation and digital twin dashboard. (a) Residential building (green) and workplace (red) locations. (b) CARTA public transit routes from public GTFS. (c) Sample On-demand configuration, where each hexagon represents service zones for respective on-demand services. (d) Performance metrics for the given configuration.

ACKNOWLEDGMENT

This work is based upon the work sponsored in part by the National Science Foundation under grants 2238815 and 1952011. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. Results reported in this paper were obtained using the Chameleon testbed, which is supported by NSF.

REFERENCES

[1] S. Li, Z. Yan, and C. Wu, "Learning to delegate for large-scale vehicle routing," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 26 198–26 211, 2021.

[2] B. Xu, Y. Wang, Z. Wang, H. Jia, and Z. Lu, "Hierarchically and cooperatively learning traffic signal control," in *AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 669–677.

[3] W. Du, J. Ye, J. Gu, J. Li, H. Wei, and G. Wang, "Safelight: A reinforcement learning method toward collision-free traffic signal control," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 801–14 810.

[4] X. Xu, Z. Zheng, Z. Hu, K. Feng, and W. Ma, "A unified dataset for the city-scale traffic assignment model in 20 us cities," *Scientific Data*, vol. 11, no. 1, p. 325, 2024.

[5] W. Wang, W. Jiang, B. Zhang, Q. Zhu, and C. Liao, "A real network environment dataset for traffic analysis," *Scientific Data*, vol. 12, no. 1, pp. 1–12, 2025.

[6] D. Perlman, K. Tufte, L. Flint, and T. Reel, "Emerging data science for transit: Market scan and feasibility analysis," U.S. Department of Transportation, Federal Transit Administration, Washington, D.C., Tech. Rep. FTA Report No. 0218, June 2022, prepared by the John A. Volpe National Transportation Systems Center. [Online]. Available: https://www.transit.dot.gov/sites/fta.dot.gov/files/2022-06/FTA-Report-No-0218.pdf

[7] National Research Council and National Academies of Sciences, Engineering, and Medicine, *Using the American Community Survey: Benefits and Challenges*, C. F. Citro and G. Kalton, Eds. Washington, DC: The National Academies Press, 2007.

[8] J. M. Abowd, J. Haltiwanger, and J. Lane, "Integrated longitudinal employer-employee data for the united states," *American Economic Review*, vol. 94, no. 2, pp. 224–229, 2004.

[9] P. Mooney, M. Minghini *et al.*, "A review of openstreetmap data," *Mapping and the citizen sensor*, pp. 37–59, 2017.

[10] M. P. Heris, N. L. Foks, K. J. Bagstad, A. Troy, and Z. H. Ancona, "A rasterized building footprint dataset for the united states," *Scientific Data*, vol. 7, no. 1, p. 207, 2020.

[11] R. Kitchin, *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[12] P. Thakuriah, N. Y. Tilahun, and M. Zellner, *Introduction to seeing cities through big data: Research, methods and applications in urban informatics*. Springer, 2017.

[13] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2017.

[14] X. Liu, Y. Xia, Y. Liang, J. Hu, Y. Wang, L. Bai, C. Huang, Z. Liu, B. Hooi, and R. Zimmermann, "Largest: A benchmark dataset for large-scale traffic forecasting," in *Advances in Neural Information Processing Systems*, 2023.

[15] Q. Li, Z. M. Peng, L. Feng, Z. Liu, C. Duan, W. Mo, and B. Zhou, "Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 3894–3920, 2023.

[16] T. Kashiyama, Y. Pang, and Y. Sekimoto, "Open pflow: Creation and evaluation of an open dataset for typical people mass movement in urban areas," *Transportation Research Part C: Emerging Technologies*, vol. 85, 2017.

[17] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," *IEEE Transactions on Mobile Computing*, vol. 13, no. 5, pp. 1061–1075, 2013.

[18] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.

[19] N. E. Lownes and R. B. Machemehl, "Vissim: a multi-parameter sensitivity analysis," in *Proceedings of the 38th Conference on Winter Simulation*, ser. WSC '06. Winter Simulation Conference, 2006, p. 1406–1413.

[20] E. Cipriani, S. Gori, and M. Petrelli, "Transit network design: A procedure and an application to a large urban area," *Transportation Research Part C: Emerging Technologies*, vol. 20, no. 1, pp. 3–14, 2012.

[21] J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, and D. Rus, "On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment," *Proceedings of the National Academy of Sciences*, vol. 114, no. 3, 2017.

[22] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering spatial patterns in origin-destination mobility data," *Transactions in GIS*, vol. 16, no. 3, pp. 411–429, 2012.

[23] V. Guihaire and J.-K. Hao, "Transit network design and scheduling: A global review," *Transportation Research Part A: Policy and Practice*, vol. 42, no. 10, 2008.

[24] S. Liyanage and H. Dia, "An agent-based simulation approach for evaluating the performance of on-demand bus services," *Sustainability*, vol. 12, no. 10, p. 4117, 2020.

[25] R. Sen, T. Tran, S. Khaleghian, P. Pugliese, M. Sartipi, H. Neema, and A. Dubey, "Bte-sim: Fast simulation environment for public transportation," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 2886–2894.

[26] United States Census Bureau, "Longitudinal Employer-Household Dynamics - Origin-Destination Employment Statistics," 2017.

[27] P. R. Stopher and S. P. Greaves, "Household travel surveys: Where are we going?" *Transportation Research Part A: Policy and Practice*, vol. 41, no. 5, 2007.

[28] S. Peer, J. Knockaert, P. Koster, and E. T. Verhoef, "Over-reporting vs. overreacting: Commuters' perceptions of travel times," *Transportation Research Part A: Policy and Practice*, vol. 69, pp. 476–494, 2014.