



Comparing real-world and simulated mobility data

A path-based, higher-order approach

IC2S2 '26
July 30th, 2026
Burlington, VT

Simulating Individual Human Mobility

Perennial Challenge

- ▶ Collecting individual mobility trace (i.e., GPS) data is often **difficult**, **expensive**, and even **risky**
- ▶ Dream: What if we could **simulate high-quality, realistic individual mobility** by combining collective mobility patterns with agent-based simulations?

Advantages of Simulated Data

- ▶ **Obviates need to collect detailed individual data**, avoiding privacy, logistics, and cost issues
- ▶ **Enables transferability** to regions without GPS data
- ▶ **Uses existing mobility data** collected by governments, transportation firms, and NGOs
- ▶ **Facilitates simulation of artificial interventions** to help guide policymaking



Key Features

- | | |
|--|--|
|  Agent-Based Traffic Simulation
Simulate a large number of individual travelers in your transport model. |  Private and Public Traffic
Simulate various transport modes, including car, transit, walk, bike, dtt, ... |
|  Fast and Dynamic
MATSim simulates whole days second-by-second for fully dynamic results. |  Supports large scenarios
MATSim can simulate large metropolitan areas in high level of details. |
|  Active Development
MATSim continues to evolve and regularly gains new functionality. |  Versatile Analyses and Simulation Output
Use preprocessed output or extract your own KPIs from the detailed simulation output. |
|  Open Source
Adapt the code to your needs. |  Modular Approach
Extend or replace functionality and algorithms based on your needs. |

Making the dream a reality

- ▶ We need methods to **evaluate whether simulated datasets are reliable surrogates** for real human mobility.

Today: Path-based Mobility Analysis

Our Approach

Compare real-world and simulated individual mobility data from a **path-based higher-order** perspective.

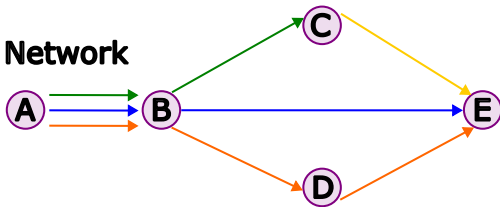
Guiding Questions

- ▶ Do simulated mobility paths **capture higher-order patterns** observed in real mobility paths?
- ▶ How well does a simulated dataset **align with observed mobility** in the same region?

Path Data

人	C	E		
人	A	B	C	
人	A	B	E	
人	A	B	D	E

Network



Case Study: Île-de-France

The NetMob 2025 Data Challenge is organized with the support of the following institutions and research projects:



NetMob 2025 Data Challenge

- ▶ Mobility trajectories for 3,320 people (representative sample) living in Île-de-France (Chasse *et al.* 2025)
- ▶ Participants carried dedicated GPS devices for 1 week and logged trip information
- ▶ Each trip is a path through the transportation network

Open Île-de-France MATSim Simulation

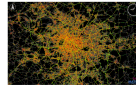
- ▶ Synthetic population of Île-de-France built from open census and regulatory data (Hörl & Balać, 2021)
- ▶ MATSim simulation creates synthetic mobility paths through the transportation network

OpenStreetMaps Network

We use the OpenStreetMaps (OSM) network as the underlying transportation network in both cases.



An open synthetic population of Île-de-France



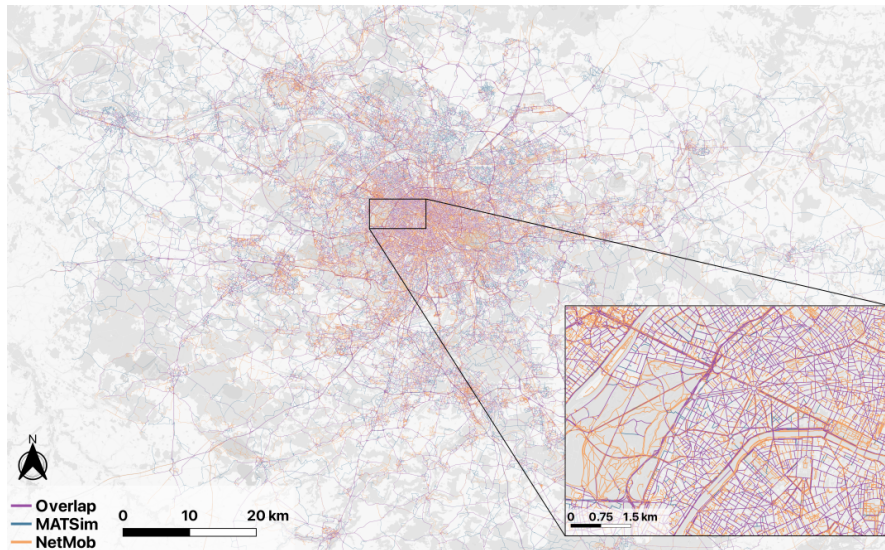
This repository contains the code to create an open data synthetic population of the Île-de-France region around in Paris and other regions in France.



Path Data

Mode	NetMob Mode Name	NetMob	MatSim	Combined Total
Vehicle	PRIV_CAR_DRIVER	20,915		
	PRIV_CAR_PASSENGER	2,952		
	LIGHT_COMM_VEHICLE	44		
	TAXI	195		
	TWO_WHEELER	530		
	Total Vehicle	24,636	22,861	47,253
Bicycle	BIKE	3,693		
	ELECT_BIKE	1,685		
	ELECT_SCOOTER	348		
	Total Bicycle	5,726	1,223	6,902
Walking	WALKING	17,571	18,685	35,260
Public Transportation	BUS	2,728	10,947	13,648
Total	–	50,661	53,716	103,063

Mapped Data



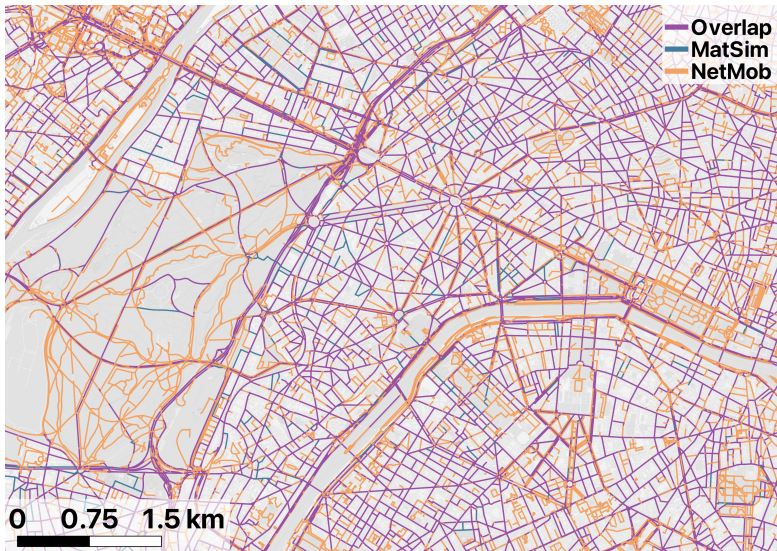
Mapped Data

Observed paths are more diverse

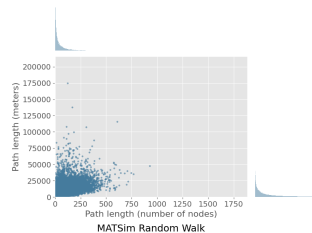
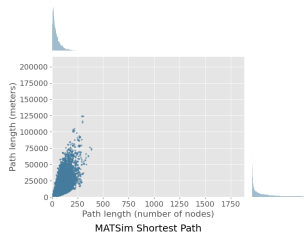
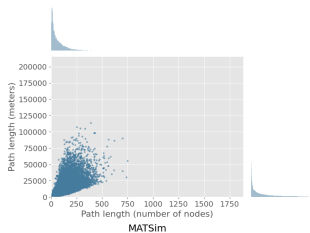
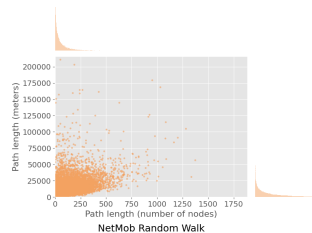
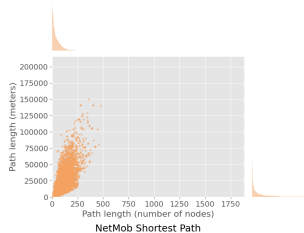
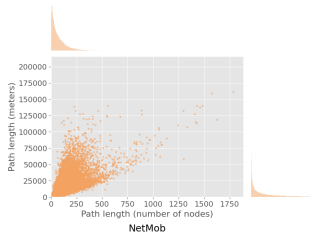
- ▶ Region between Bois de Boulogne in the west to the Louvre in the east
- ▶ NetMob coverage **much more diverse in pedestrian areas**

Why?

- ▶ MATSim provides only origin and destination for walking and cycling, we impute shortest paths



Path Length Distributions



Observed Data

Shortest Path Baseline

Random Walk Baseline

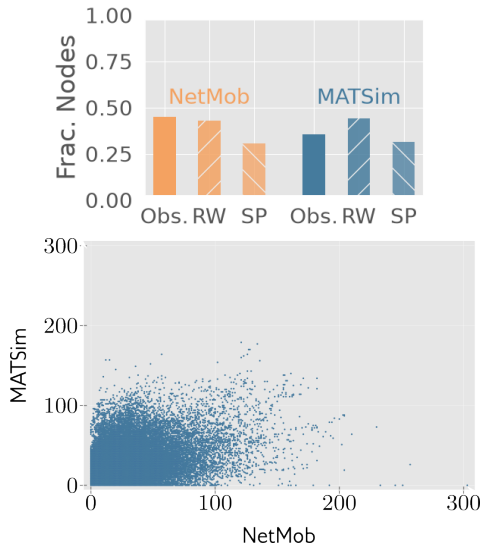
Network Coverage & Visitation Alignment

Node visits

- ▶ What **proportion of nodes** are visited by paths from each dataset?
- ▶ Are **node visit frequencies** aligned?

Takeaways

- ▶ Fewer than **half of nodes in the network visited**
- ▶ Visitation **frequencies correlate, but with some outliers**



Beyond Nodes: Subpath Alignment

Subpath overlap

Given path $p = \langle u_0, u_1, \dots, u_\ell \rangle$, slide window of length k edges for $k \in 0, \dots, 100$.

Example: $\langle a, b, c, d, e \rangle$ for $k = 2$ edges results in:

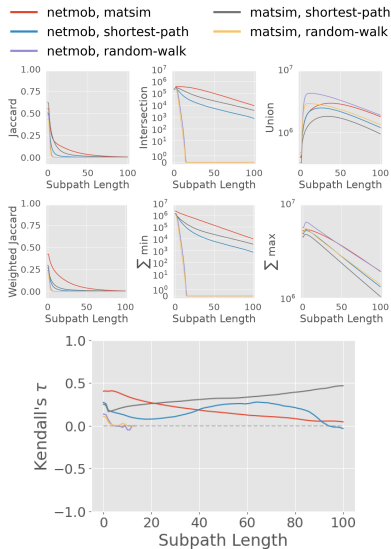
$$\{\langle a, b, c \rangle, \langle b, c, d \rangle, \langle c, d, e \rangle\}.$$

Subpath Comparisons

- ▶ Unweighted: Intersection over union of subpath sets
- ▶ Weighted: Sum of min/max observed visitation ratio
- ▶ Kendall's τ of subpaths ranked by frequency

Takeaway

- ▶ Shared signal between real and simulated mobility (sub)paths weakens with length



Thank You!

Takeaways

- ▶ Positive: Broad overall alignment; node/edge/path visitations have some correlation; same estimated optimal memory length.
- ▶ Less positive: Simulations miss extreme paths; pedestrian modes poorly modeled; simulated paths more predictable.

Contact

Dr. Timothy LaRock
tlarock@princeton.edu
<https://tlarock.github.io>



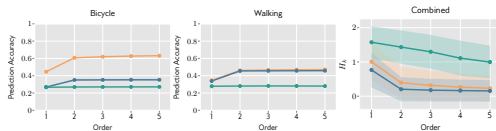
OpenStreetMap

The NetMob 2025 Data Challenge is organized with the support of the following institutions and research projects:



Also in the pre-print...

- ▶ Many details on data pre-processing
- ▶ Next-step prediction & entropy/perplexity analysis



Higher-order Network Analysis of Human Mobility Data

Timothy LaRock, Chen Zhang, Jürgen Hackl

Subjects: **Computational Engineering, Finance, and Science (cs.CE)**

Cite as: [arXiv:2606.00733](https://arxiv.org/abs/2606.00733) [cs.CE]

(or [arXiv:2606.00733v1](https://arxiv.org/abs/2606.00733v1) [cs.CE] for this version)

<https://doi.org/10.48550/arXiv.2606.00733> ⓘ

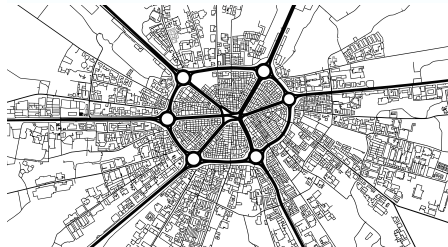
Human Mobility

Fundamental Questions

- ▶ How do people **move through the world**?
- ▶ What factors **explain or influence** mobility behavior?
- ▶ How does human mobility **evolve over time**?
- ▶ Can we make **accurate predictions** to help **guide decision making**?

Application areas

- ▶ planning and design
- ▶ transportation engineering
- ▶ public health and epidemiology
- ▶ point-of-interest recommendation
- ▶ ...and many more



Connecting Mobility and Infrastructure Networks

Why infrastructure?

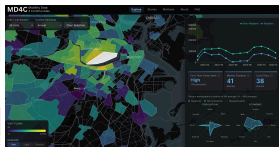
- ▶ Transportation infrastructure influences (and is influenced by) mobility patterns
- ▶ Policymaking requires understanding how people actually use physical infrastructure

Geospatial aggregation breaks the link

- ▶ Mobility studies often aggregate based on geospatial tessellation (grids, census blocks, H3)
- ▶ Simplifies network construction, but **loses connection to infrastructure**

Our work: mobility dynamics + infrastructure networks

Integrate **individual mobility** dynamics and **infrastructure topology** using **higher-order network analysis**



Challenges in Individual Mobility Research

Data Collection

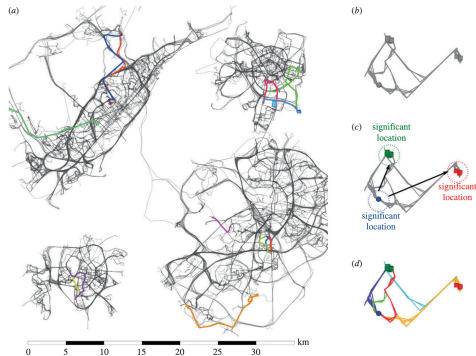
- ▶ **Difficult to collect** good data due to costs, logistics, and privacy considerations
- ▶ Even good data is inherently **large, noisy, and heterogeneous**

Data Integration

- ▶ Requires integrating **geospatial, network, and administrative/economic/environmental** datasets
- ▶ Messy and diverse data needs pre-processing, adding uncertainty

Our Question

Can simulations help overcome these challenges?



Random Walks on Graphs

Discrete time random walk

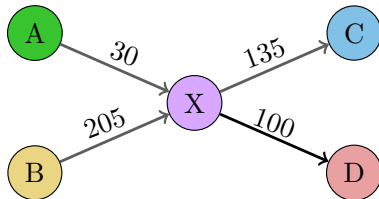
- ▶ Random walker placed on node u at time $t = t_0$
- ▶ Let η_u denote the out-neighbors of u
- ▶ At $t + 1$, the walker moves to a node $v \in \eta_u$ with the following probability:

$$\Pr(v|u) = \frac{w_{uv}}{\sum_{x \in \eta_u} w_{ux}}$$

where w_{uv} , $v \in \eta_u$ is either 1 (*unweighted* or *uniform*) or the weight of the edge between u and v (*edge weighted*).

Note

- ▶ This process is called **memoryless** (also **Markovian**) because the next node to be visited depends only on the current node



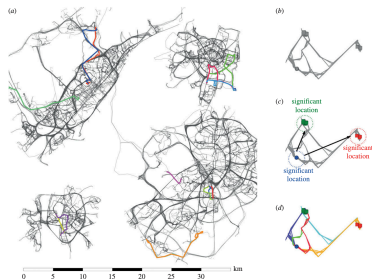
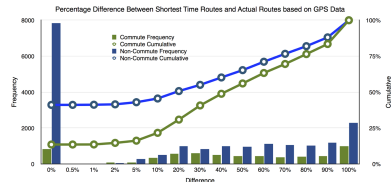
Higher-order Human Mobility

Humans are not (memoryless) random walkers

- ▶ People move with **intentionality** and **memory**
- ▶ Transportation networks impose **constraints**
- ▶ During a trip, individuals **take action** in response to both **internal goals** and the **external environment**

Humans are not shortest-path machines

- ▶ Preferences for efficiency are **contextual**, not based purely on shortest distance or travel time
- ▶ Factors like infrastructure quality and ease of use, local knowledge, urgency, and habit matter



Preprocessing (in brief)

Network simplification

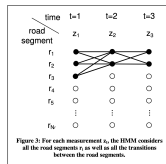
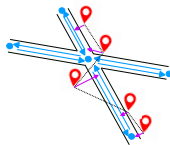
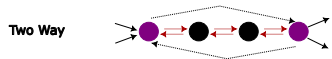
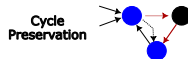
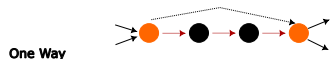
- ▶ Remove **redundant nodes** that do not represent meaningful choice points for travelers

NetMob Map matching

- ▶ Match GPS coordinates to sequences of OpenStreetMaps network edges
- ▶ Implementation: Hidden Markov Model in Valhalla open source routing engine

MATSim Path extraction

- ▶ Extract sequences of OSM nodes from MATSim output
- ▶ Interpolate walking and biking modes with shortest paths
- ▶ Map transit (bus) paths from stops back to OSM edges



Dataset Processing: Network Simplification

Network simplification

- ▶ Remove “redundant” nodes that do not represent meaningful choice points for travelers and artificially reduce navigational complexity
- ▶ Removal also reduces downstream computation

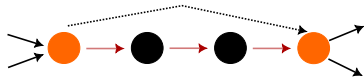
Redundant nodes

- ▶ One-way roads: Nodes with degree exactly 2 and different in/out neighborhoods
- ▶ Two-way roads: Nodes with degree exactly 4 and equivalent in/out neighborhoods

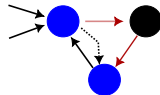
Consequences

- ▶ Nodes: Reduced from 3,016,582 to 788,018 (26%)
- ▶ Edges: Reduced from 6,655,074 to 2,197,946 (33%)
- ▶ Cycles and overall navigability preserved

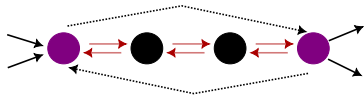
One Way



Cycle Preservation



Two Way



Dataset Processing: NetMob

Map matching

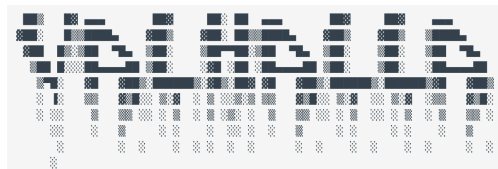
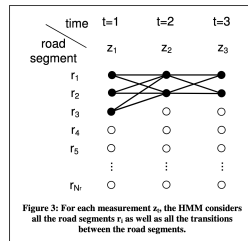
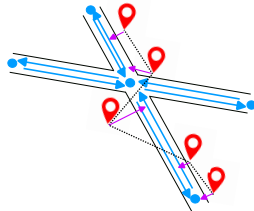
- ▶ Match sequences of GPS coordinates to OpenStreetMaps network edges
- ▶ Matched sequences of unique edges are paths

Implementation

- ▶ Hidden Markov Model approach
- ▶ Valhalla Open Source Routing Engine implementation

Post processing

- ▶ Discontinuities interpolated with shortest paths
- ▶ Redundant nodes removed from paths



Dataset Processing: MATSim

Path extraction

- ▶ Extract sequences of OSM nodes from MATSim XML output events
- ▶ Interpolate walking and biking modes (MATSim only provides origin/destination)
- ▶ Map transit (bus) paths from stops (GTFS data) back to OSM edges

Takeaways

Mobility analysis with higher-order networks

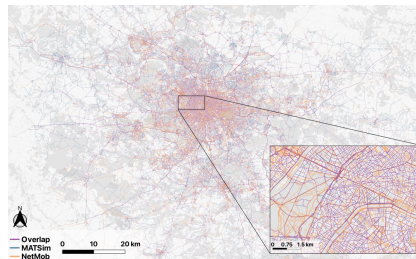
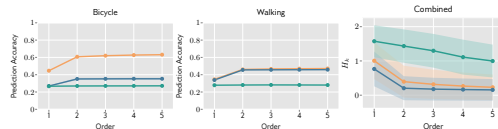
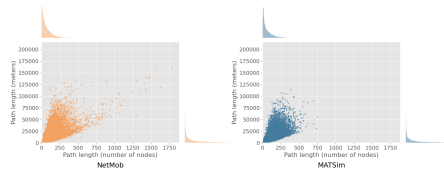
Higher-order network analysis gives us insight into the relationship between observed and simulated mobility.

Good News

- ▶ Open Île-de-France MATSim simulations and NetMob 2025 Data Challenge datasets align well overall
- ▶ Similar higher-order network structural features and overall node visitation rates

Less Good News

- ▶ MATSim does not generate paths at the extreme end of the NetMob distribution
- ▶ Node visitation patterns do not align directly
- ▶ Uncertainty with pedestrian and cycling paths
- ▶ MATSim paths are more predictable and have lower expected conditional entropy than NetMob paths



Limitations & Future Work

Map matching quality

- ▶ Without ground truth, we need methods to validate map matching on new datasets
- ▶ Multi-modal trip segmentation (pre-processing step by data provider)

Further Data Sources

- ▶ NetMob Data Challenge data contains more information (trip purpose, sociodemographics)
- ▶ Simulations beyond MATSim (SUMO, existing Digital Twin models) and augmented pedestrian simulations
- ▶ Comparison with geospatial aggregation and point-of-interest analysis methods

Additional Higher-order Methods

- ▶ Statistical anomaly detection (HYPA)
- ▶ Sequential motifs
- ▶ Direct comparison of higher-order models

