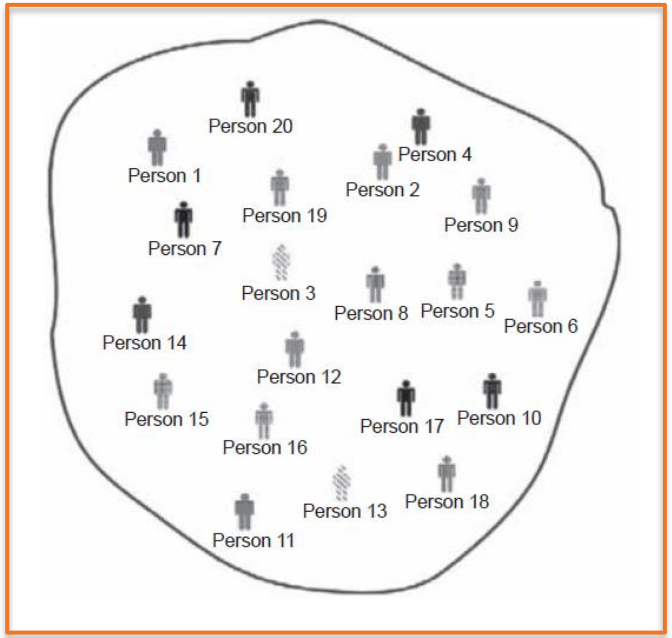
Harnessing Household Travel Survey with Passively Collected Mobility Data

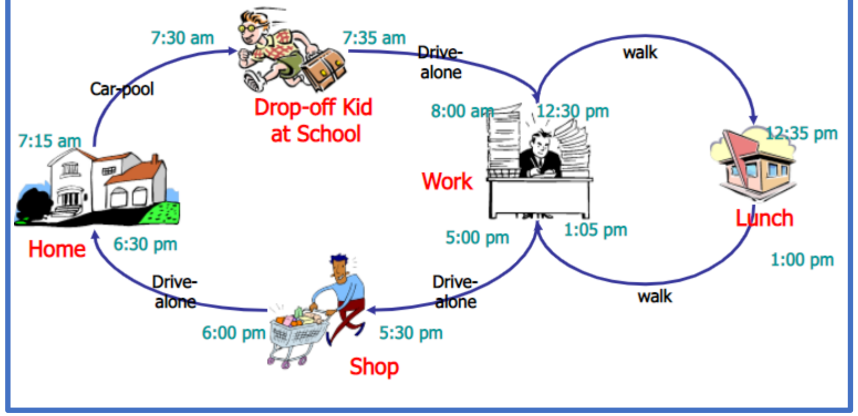
Prateek Bansal
Assistant Professor
National University of Singapore

Collaborators: Khoa Vo, Seung Woo Ham, Swapnil Mishra, Paolo Santi

Activity-based Travel Demand Model

ID	Synthetic activity schedules					Synthetic population		
Index	Location	Time	Duration	Purpose	Mode	Age	Gender	Income
1	A	09:00	10 hours	Leisure	Car	37	Male	>500
…	…	…	…	…		…	…	…
N	E	09:00	8 hours	Commute	Bus	52	Female	300<





Typical Data Sources

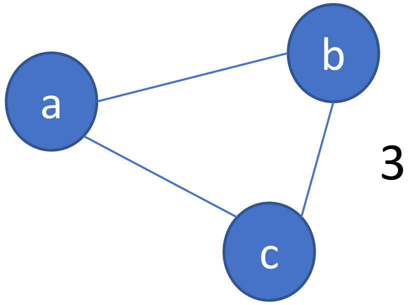
Household Travel Survey (HTS) Data

- Collect travel diary of 1-3% of the population.

Limitation of HTS dataset

- Low spatiotemporal heterogeneity due to low sampling rate (~1-3%)
- Low frequency of data collection – cannot handle shocks

Low Spatiotemporal Heterogeneity in Household Travel Survey (HTS)



3 zones

HTS (Total = 20, ~0.4%)

	a	b	c	Σ
a	3	2	2	7
b	3	4	2	9
c	2	2	0	4
Σ	8	8	4	20

Sample (N=20)

Depart ≤ 9:00

	a	b	c	Σ
a	2	2	1	5
b	1	1	1	3
c	2	0	0	2
Σ	5	3	2	10

"Zero cell" problem

Depart > 9:00

	a	b	c	Σ
a	1	0	1	2
b	2	3	1	6
c	0	2	0	2
Σ	3	5	2	10

Population (N=5400)

Depart ≤ 9:00

	a	b	c	Σ
a	300	300	400	1000
b	300	300	400	1000
c	400	200	200	800
Σ	1000	800	1000	2800

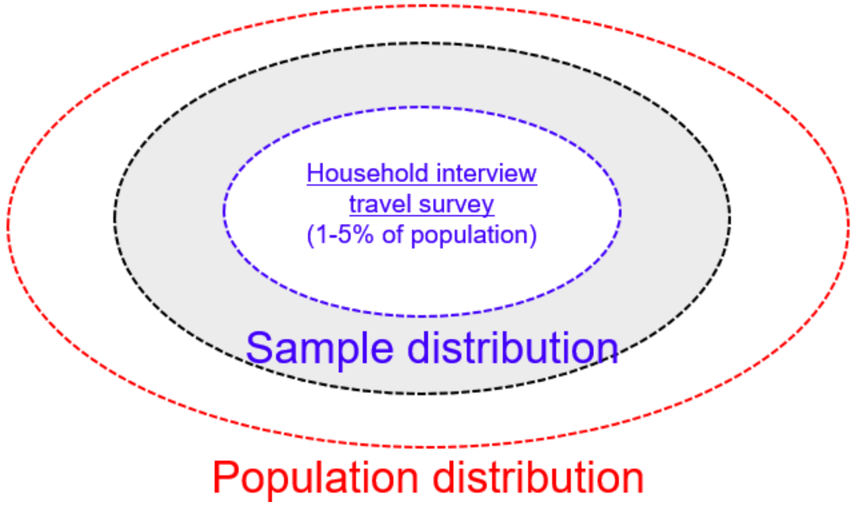
Only transit H-W trips

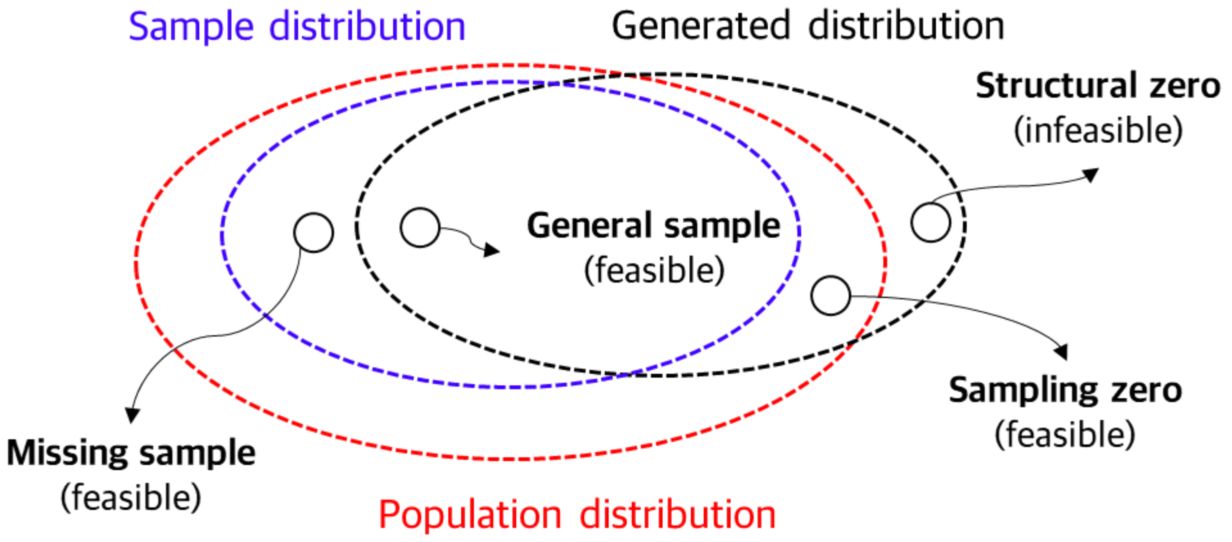
Depart > 9:00

	a	b	c	Σ
a	300	300	200	800
b	300	300	200	800
c	400	400	200	1000
Σ	1000	1000	600	2600

Potential Direction 1: Generative Modeling

- **Generate beyond sample:** Use generative models to increase the heterogeneity of HTS data



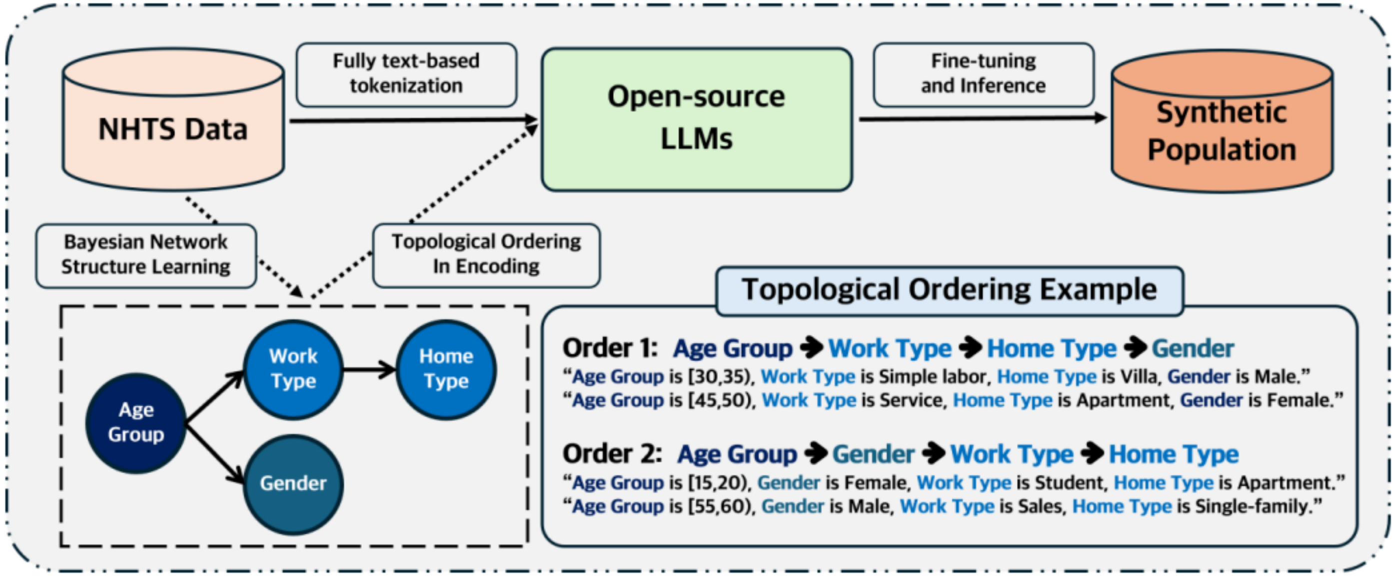


Example of structural zero: children having full-time jobs

Reference: Kim, E. J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. Transportation Research Part C: Emerging Technologies, 148, 104053.

Potential Direction 1: Generative Modeling

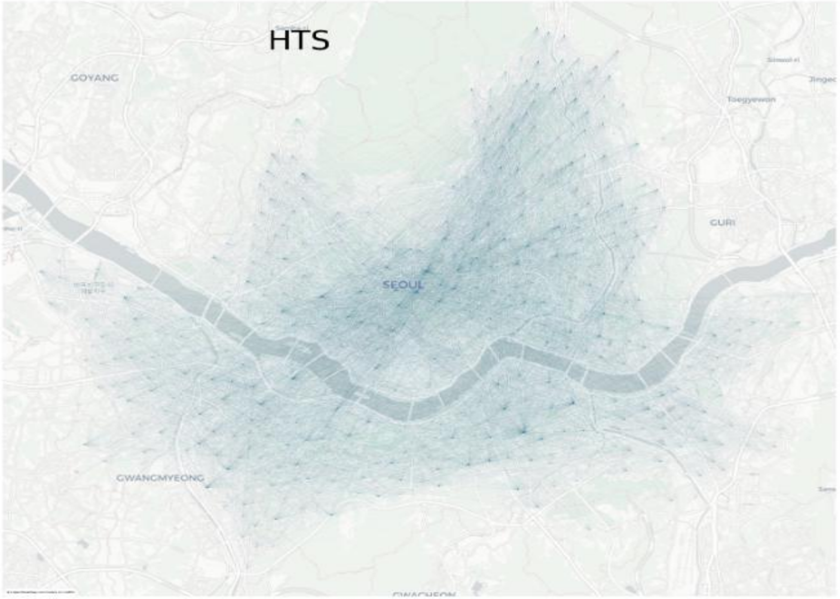
- **LLMs + Generative Models**

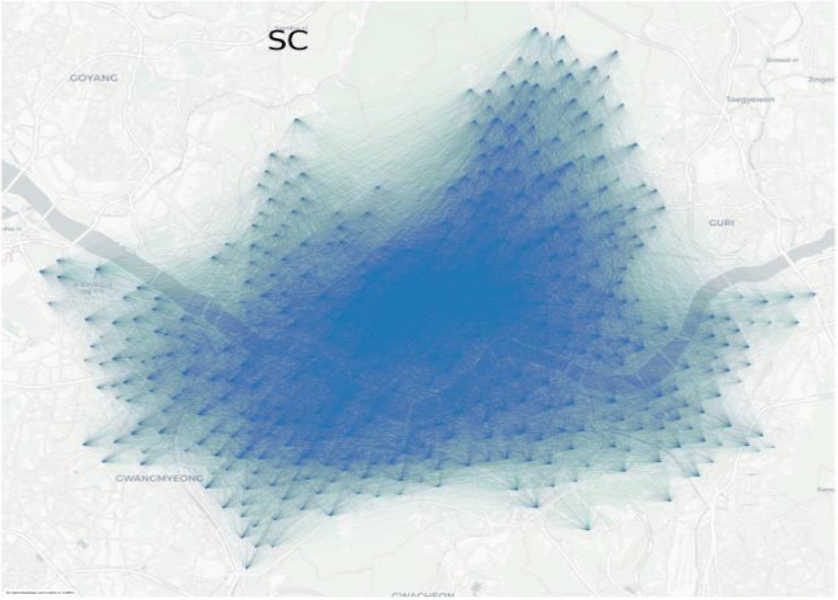


Model	Diversity		Feasibility (Precision)	Overall Quality (F1 Score)
	# of combinations	Recall		
Prototypical agent	30,837	56.4%	100.0%	72.1%
Deep generative model	263,925	80.8%	81.4%	81.1%
LLM-BN	**120,541**	76.0%	**95.3%**	84.6%

Potential Direction 2: Data Fusion

- **Data Fusion:** Fuse HTS data with passively-collected data, such as Transit Smart Card (SC) and cellular signaling Data





Data sources	Socio-demographic	Activity purpose	Travel mode	Spatial attribute	Temporal attribute
HTS	High	High	High	Low	Low
PCM	Unavailable	Unavailable	Unavailable	High	High

Prior Research on **Data Fusion**

1) A novel data fusion method to leverage passively-collected mobility data in generating spatially-heterogeneous synthetic population (2025). **Transportation Research Part B: Methodological** 191, 103128.

2) Collaborative generative adversarial networks for fusing household travel survey and smart card data to generate heterogeneous activity schedules in urban digital twins (2025). **Transportation Research Part C: Emerging Technologies** 176, 105125.

3) Harnessing household travel survey with smart card data to generate spatiotemporally heterogeneous activity schedules for transit users (Available at SSRN 4960458).

4) A data fusion framework to infer multi-modal time-dependent origin-destination travel demand matrices (Available at SSRN 5250605).

5) Scalable data fusion for generating disaggregate activity schedules (Available at SSRN 5259159).

Analytical Data Fusion of HTS with PCM data

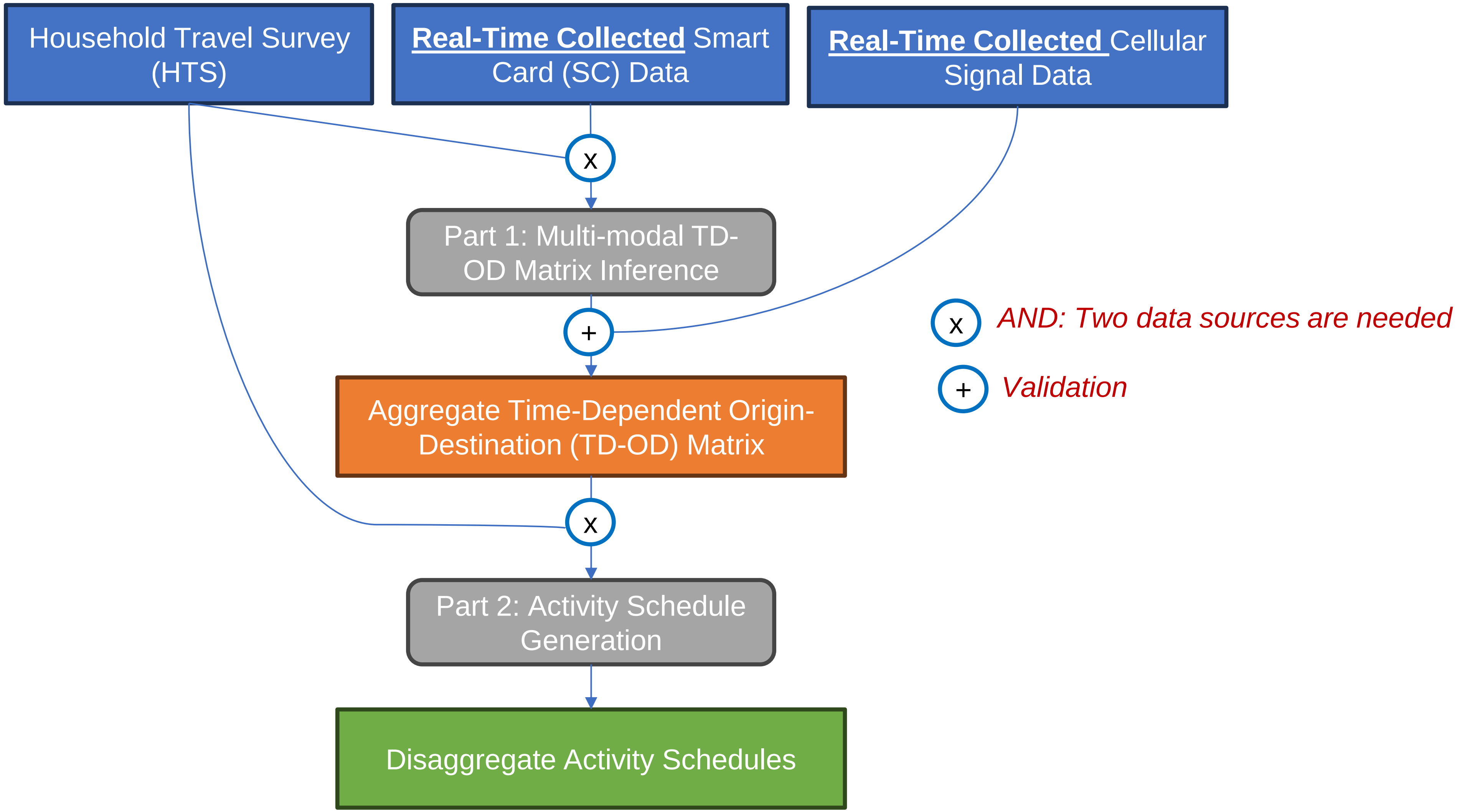
Part 1: Fusion of HTS with SC Data to Derive Time-Dependent Origin-Destination (TD-OD) Matrices for All Travel Modes

Best suited for **cities with high public transit usage** (e.g., Singapore, Seoul, Beijing, London), where detailed TD-OD matrices are lacking for non-transit modes.

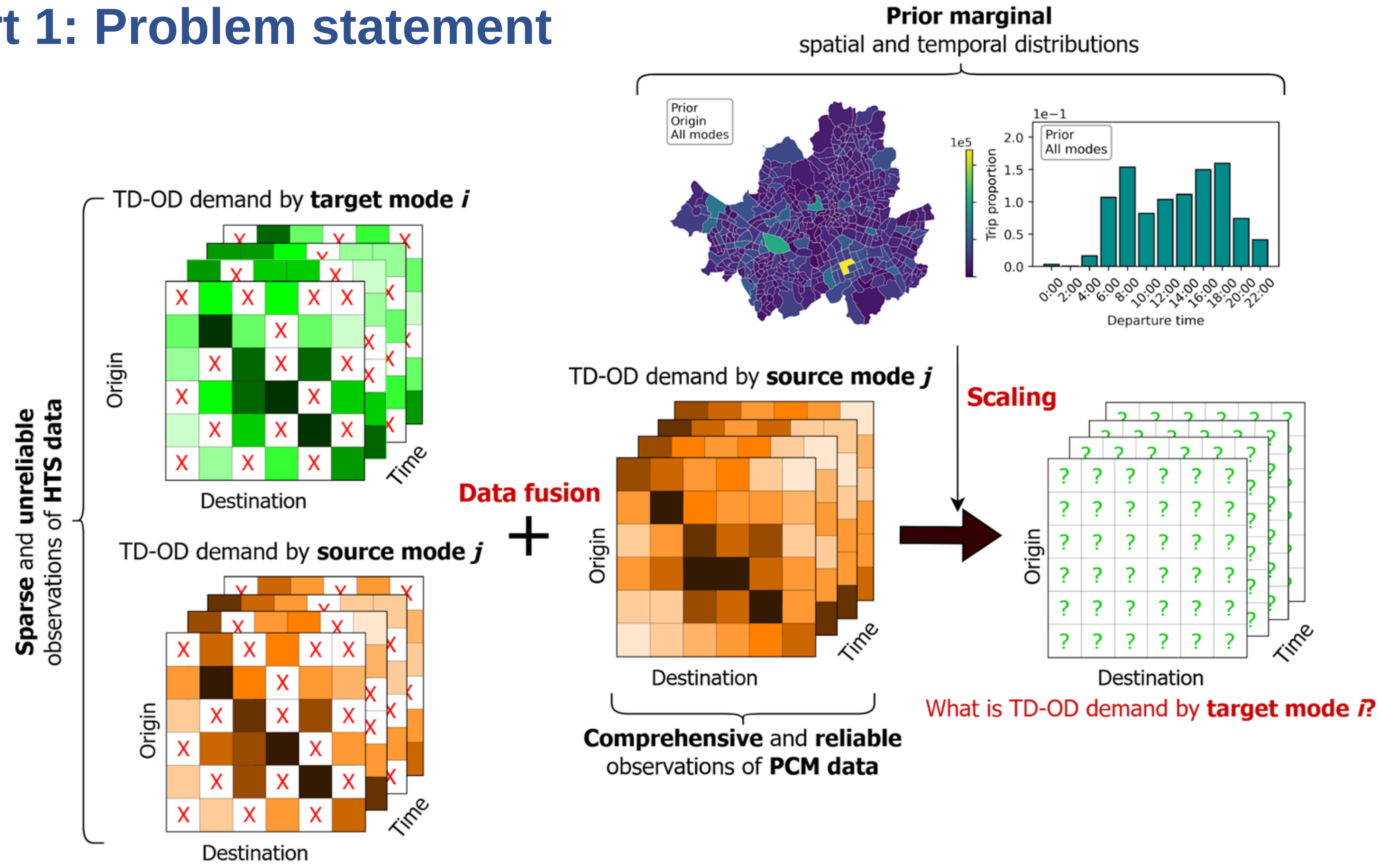
Part 2: Fusion of HTS with TD-OD Matrices to Generate Disaggregate Activity Schedules

More **globally applicable** and offers a privacy-preserving alternative by relying on aggregated mobility patterns rather than individual-level traces.

A Framework To Generate Activity Schedules Using SC and HTS



Part 1: Problem statement



Part 1: Mathematical formulation

i: target mode; w: OD; k: time

Sparse and noisy HTS observations

$$\{\ldots,(\beta_{wk}^i,\gamma_{wk}^i),\ldots\}$$

Mode share between target mode i and source mode + Sample count

Comprehensive and reliable PCM observations

$$\{\ldots,n_{wk}^{\mathrm{src}},\ldots\}$$

True travel demand for source mode

We can estimate the travel demand for target mode *i*

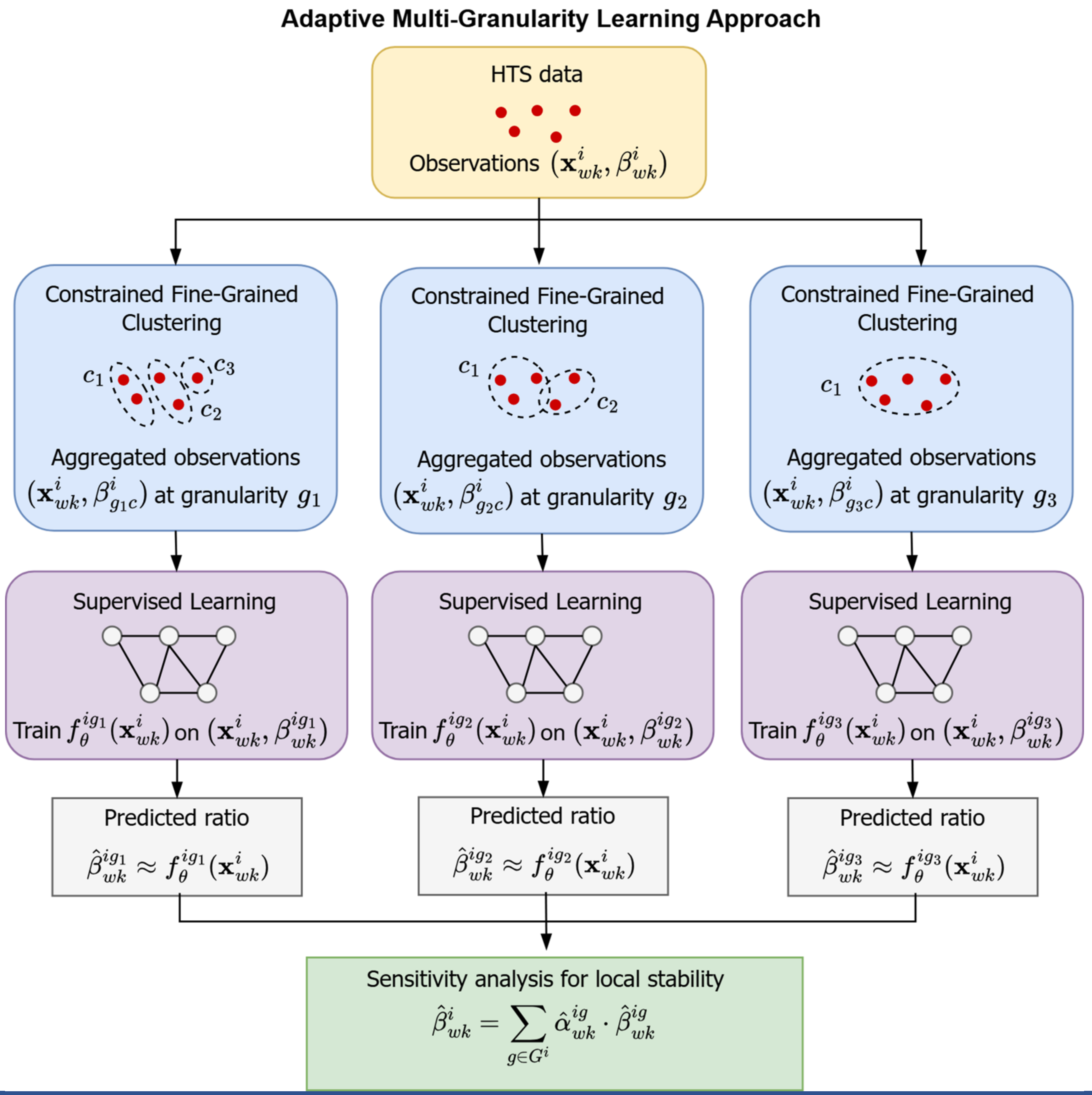
$$\hat{n}_{wk}^i = \boxed{\hat{\beta}_{wk}^i} \cdot n_{wk}^{\mathrm{src}}.$$

Finding the mode share between target mode i and source mode $\hat{\beta}_{wk}^i$

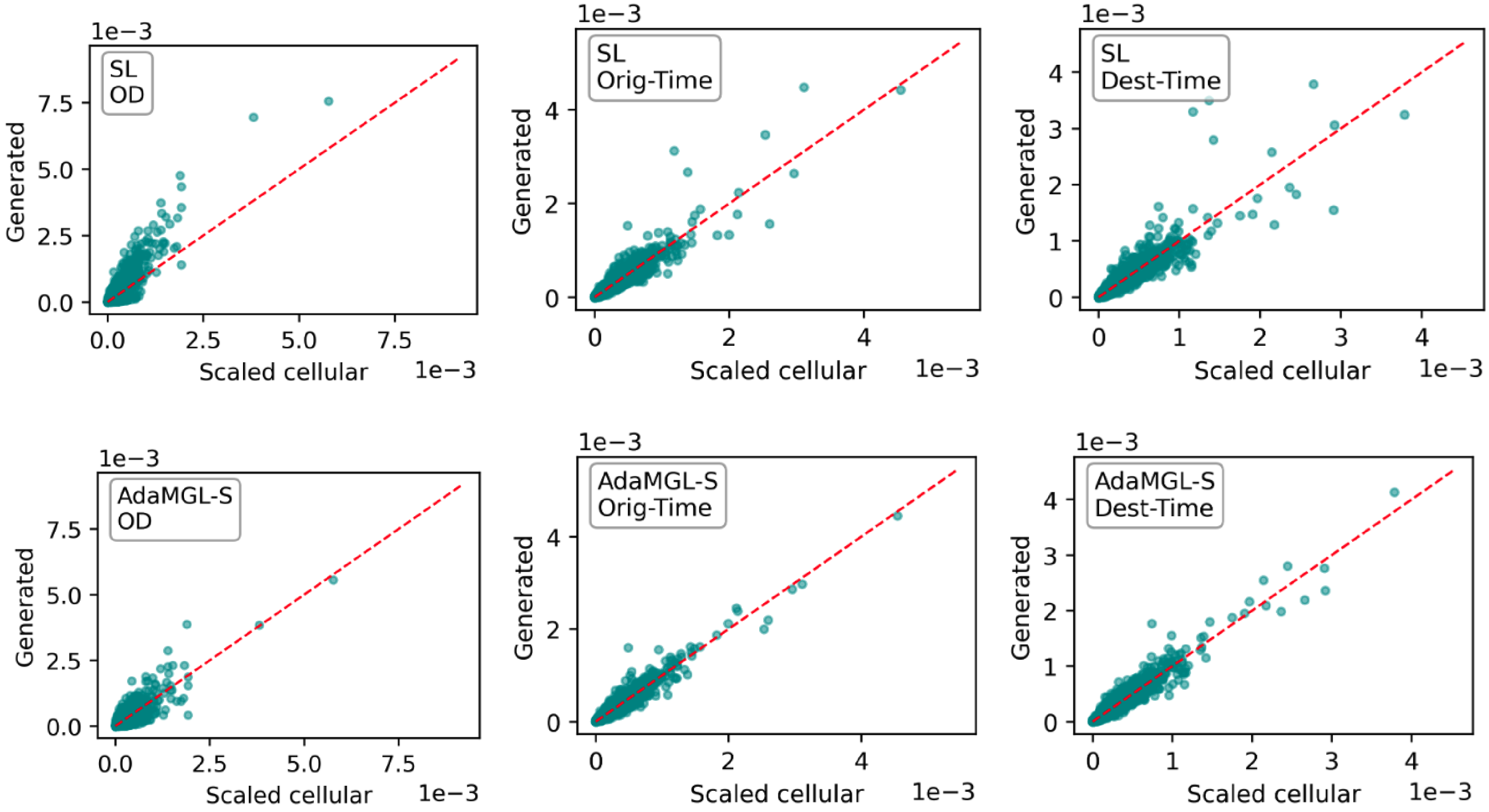
We can derive the estimated travel demand for across all modes

$$\hat{n}_{wk} = \left(1 + \sum_{i=1}^{I} \hat{\beta}_{wk}^i\right) \cdot n_{wk}^{\mathrm{src}}$$

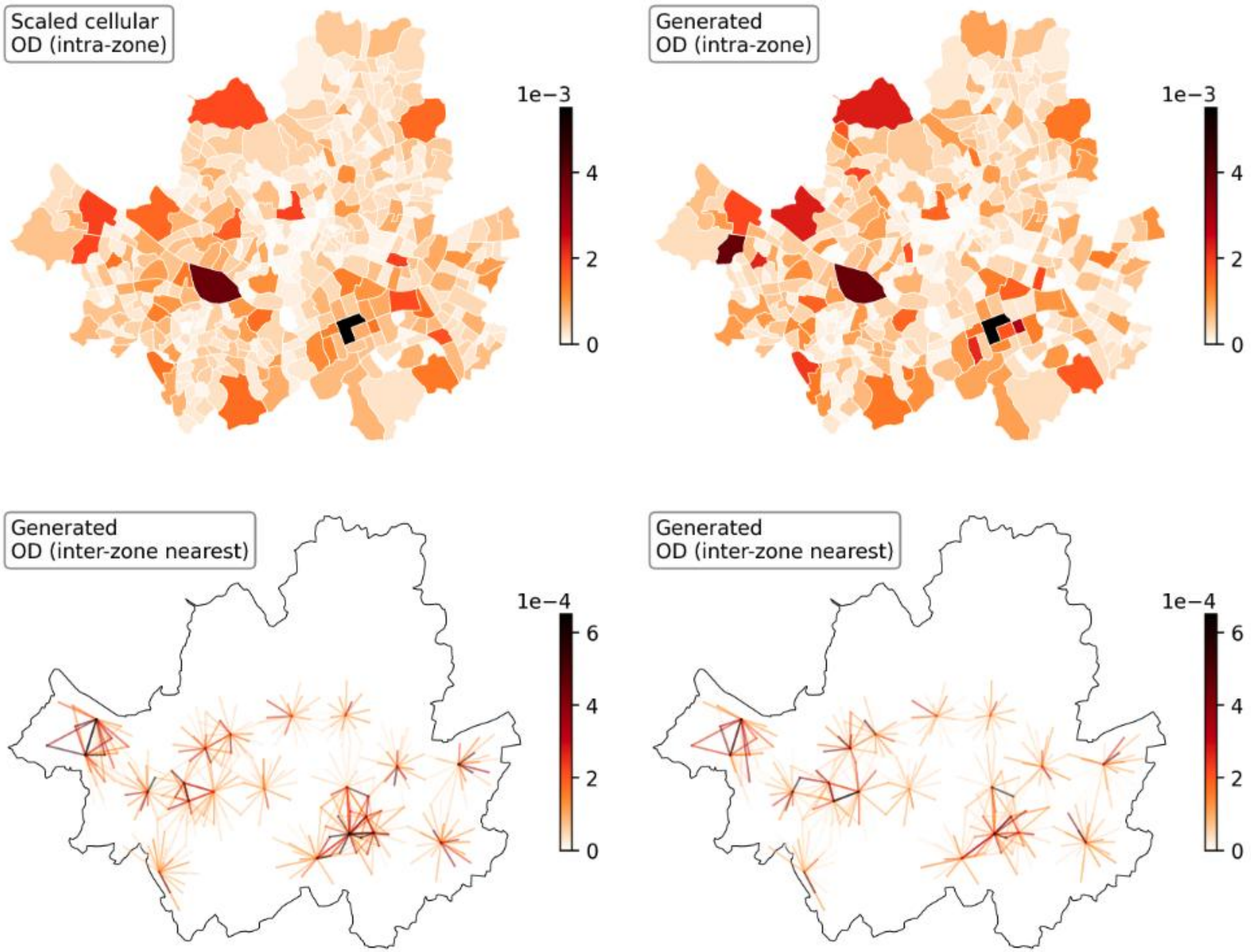
Part 1: Adaptive multi-granularity learning (AdaMGL) approach



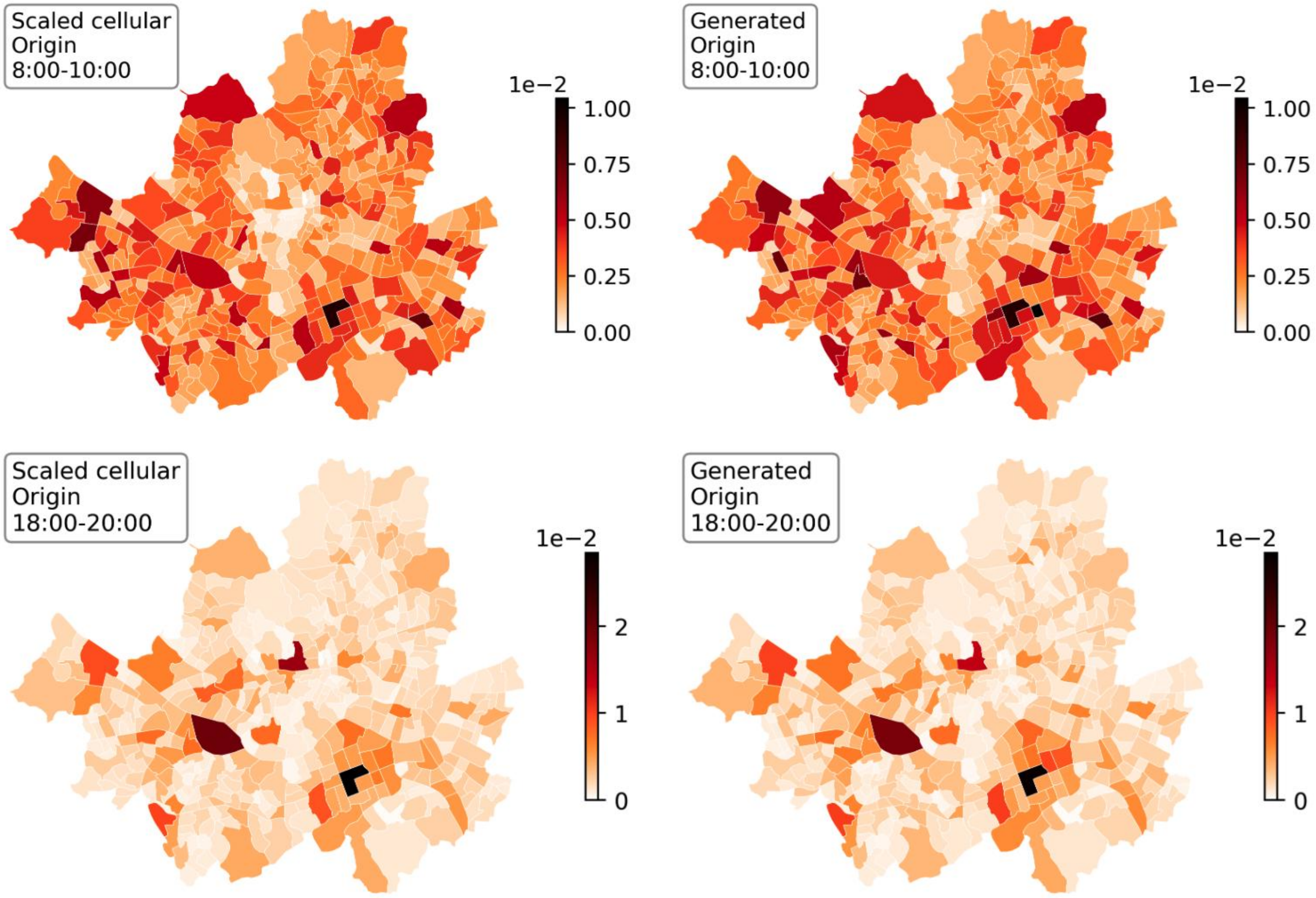
Part 1: Validation (Seoul)



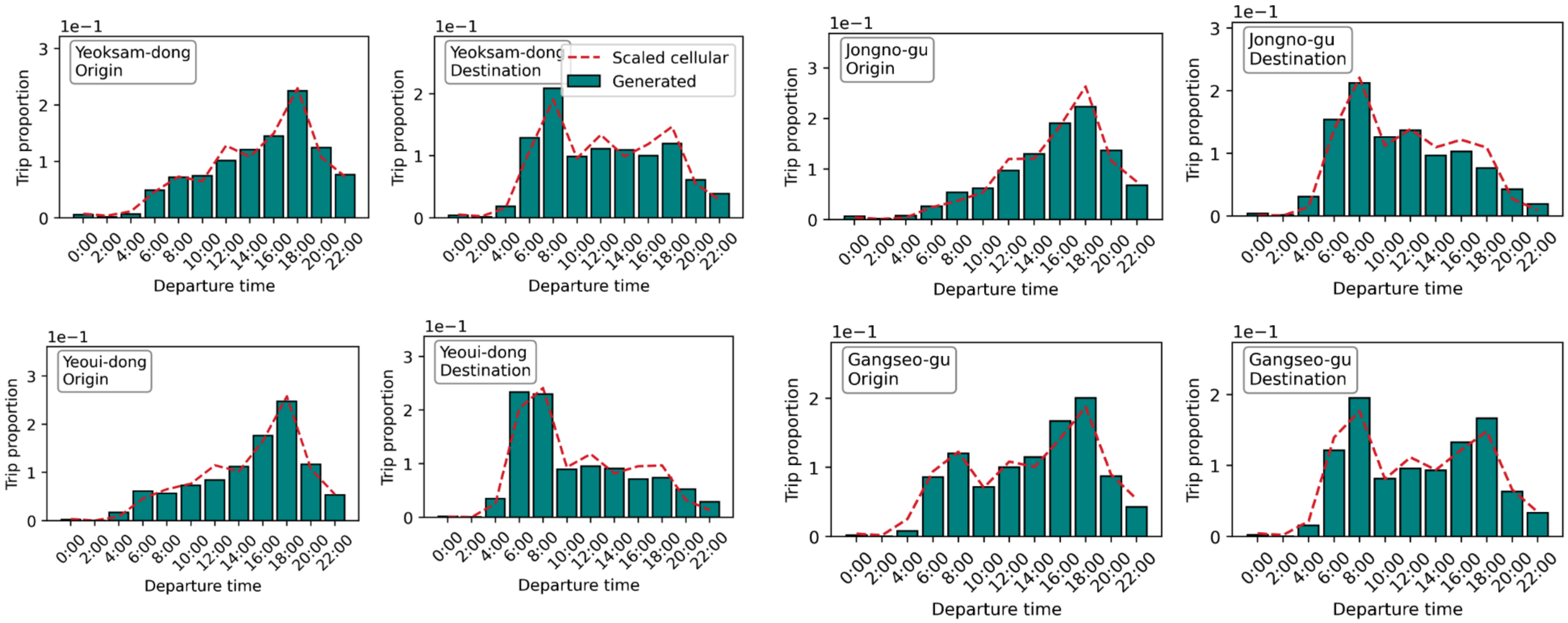
Part 1: Validation



Part 1: Validation



Part 1: Validation



Part 2: Potential of TD-OD matrix: Motivation

Time-Dependent Origin-Destination (TD-OD) matrices

- Summarize mobility as **trip volumes between locations** over time.
- **Easily obtained** from PCM data.
- **Preserves privacy** as it is an aggregated form of mobility data.
- Covers **higher proportion** of the population.

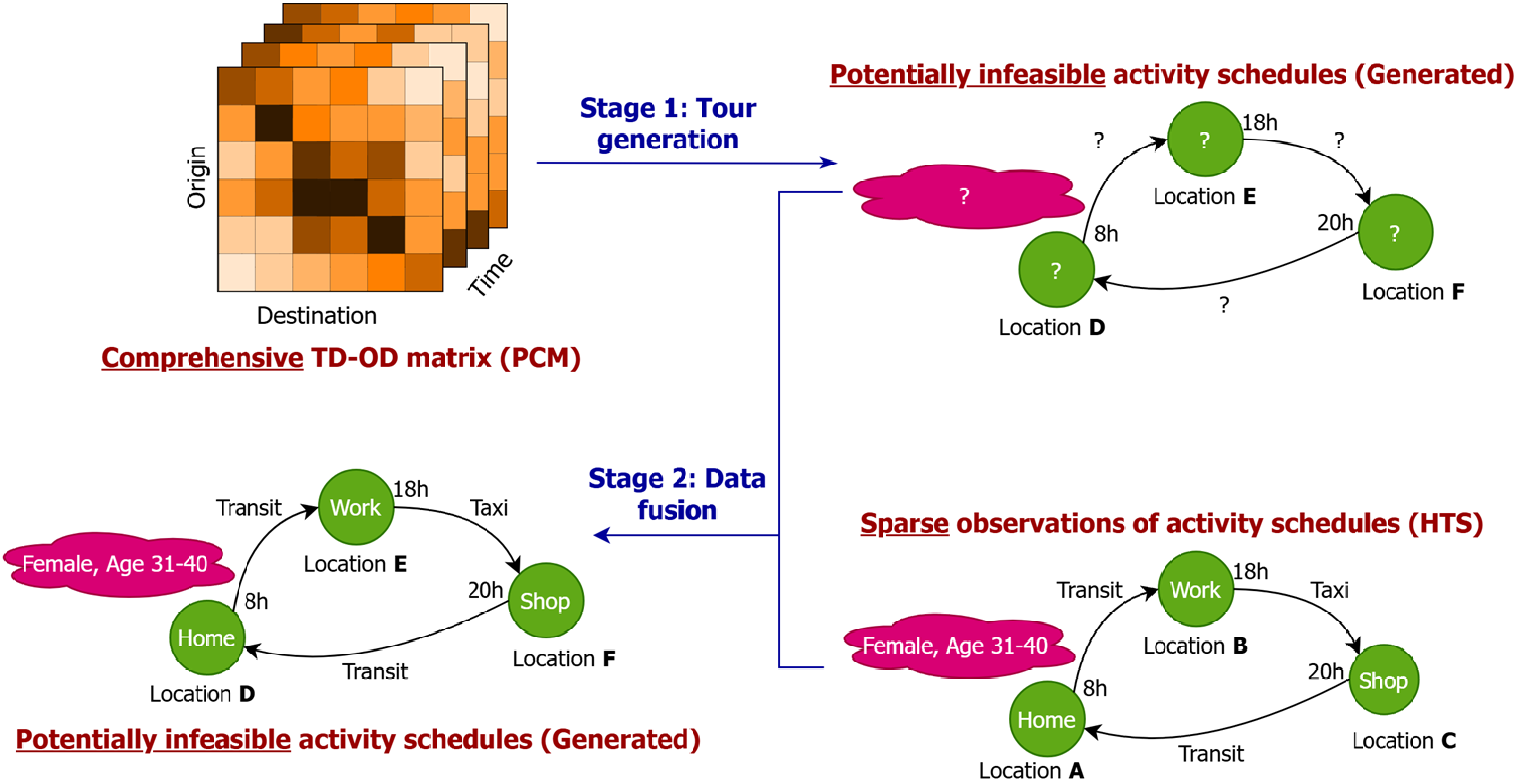
Limitations:

- Lack of **trip interdependencies**.
- Absence of **sociodemographic and trip-chain attribute details**.

Balancing data utility and privacy is key for future mobility research.

Solution: Enhancing TD-OD Matrix with HTS data

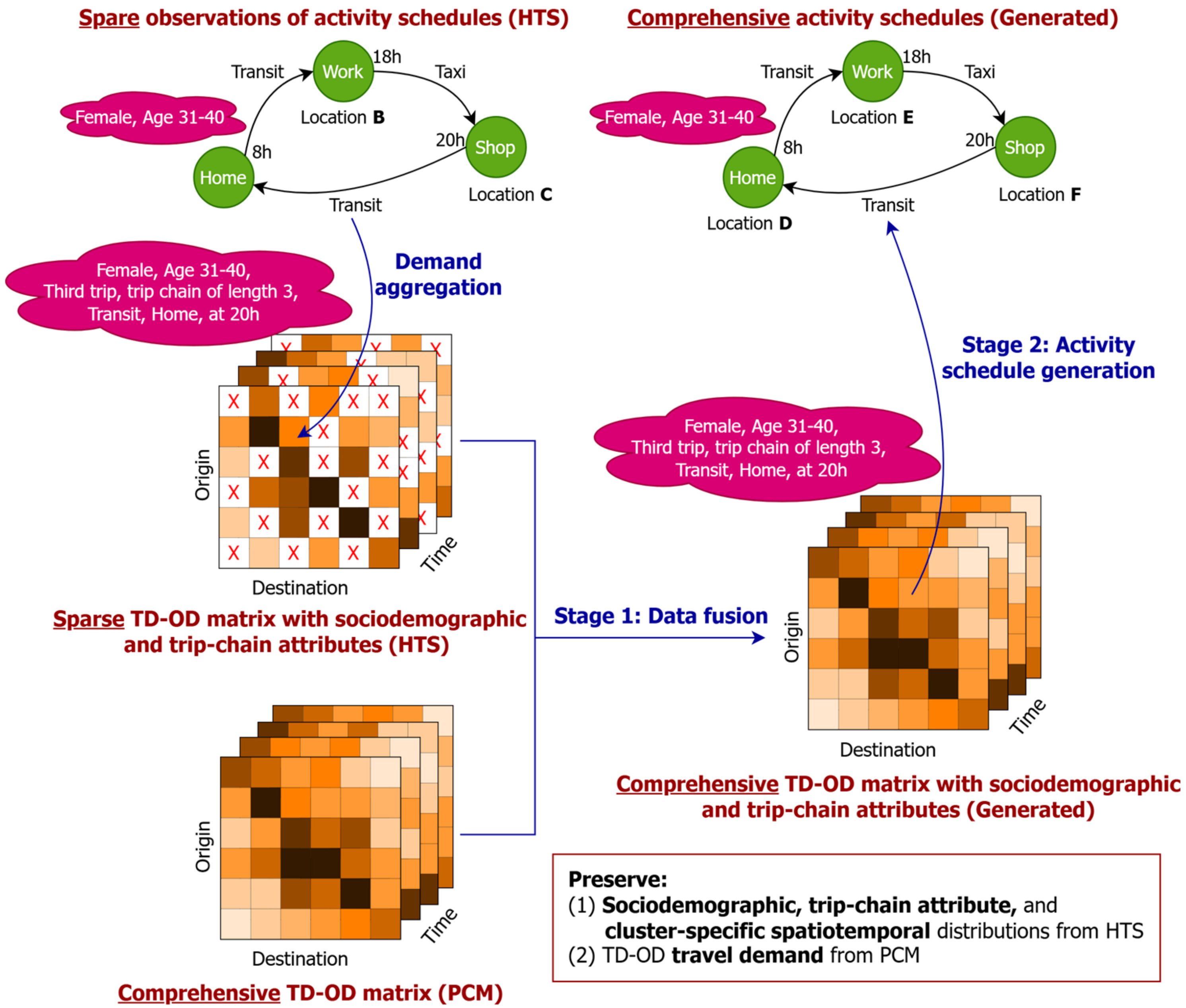
Part 2: Conventional approach



Part 2: Proposed approach

Conventional:
First tour generation, then data fusion

Proposed:
First data fusion, then activity schedule generation



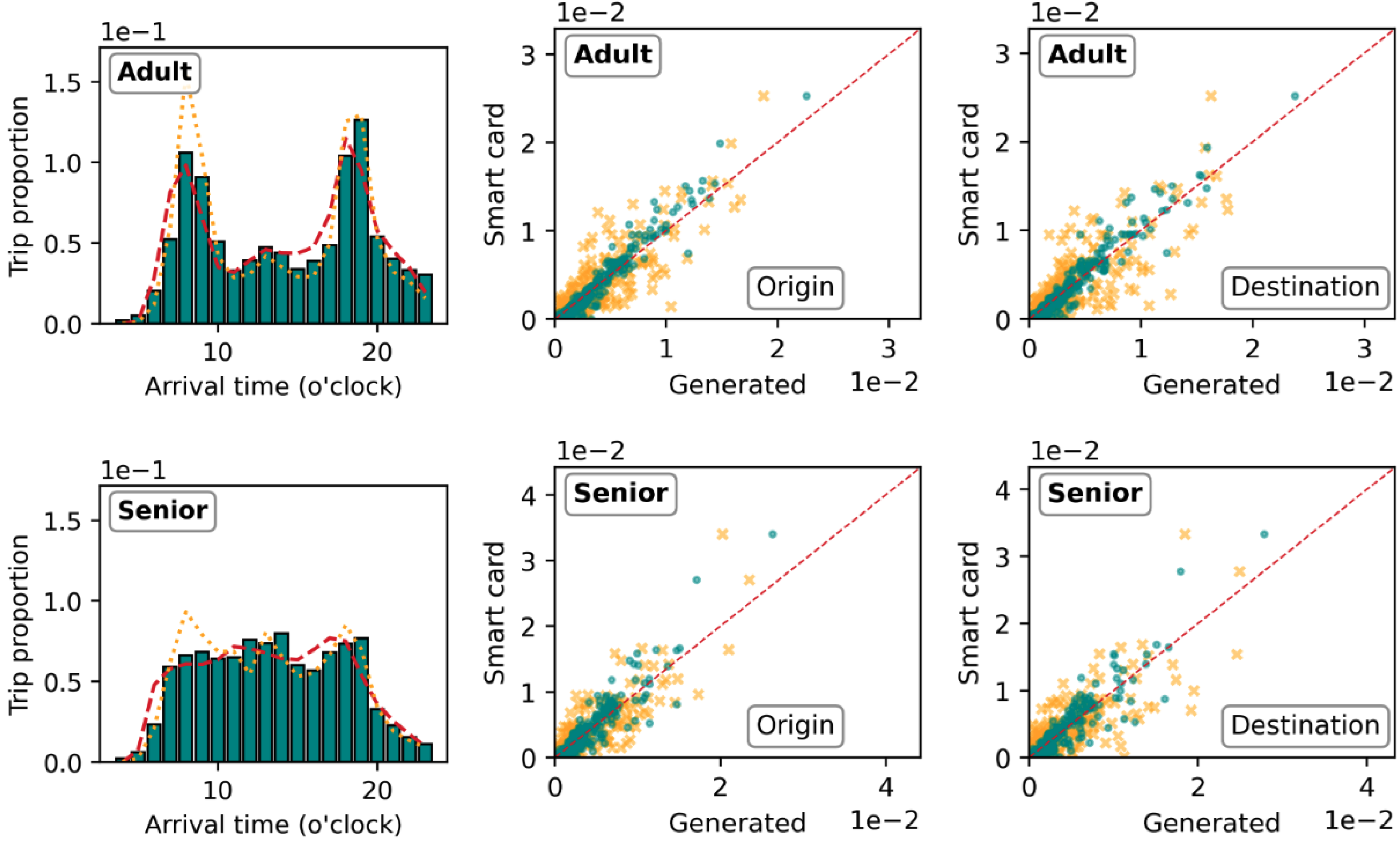
Part 2: Proposed Approach's Benefits

- **Feasibility**: The activity schedule generation in Stage 2 is be improved due to incorporation of socio-demographics and trip-chain attributes (i.e., trip-chain length and order) when determining the next activity location, start time and duration.

- **Distribution preservation**: The data fusion in Stage 1 minimizes distance from joint distribution of all attributes in HTS and preserves spatiotemporal distribution of TD-OD from PCM data.

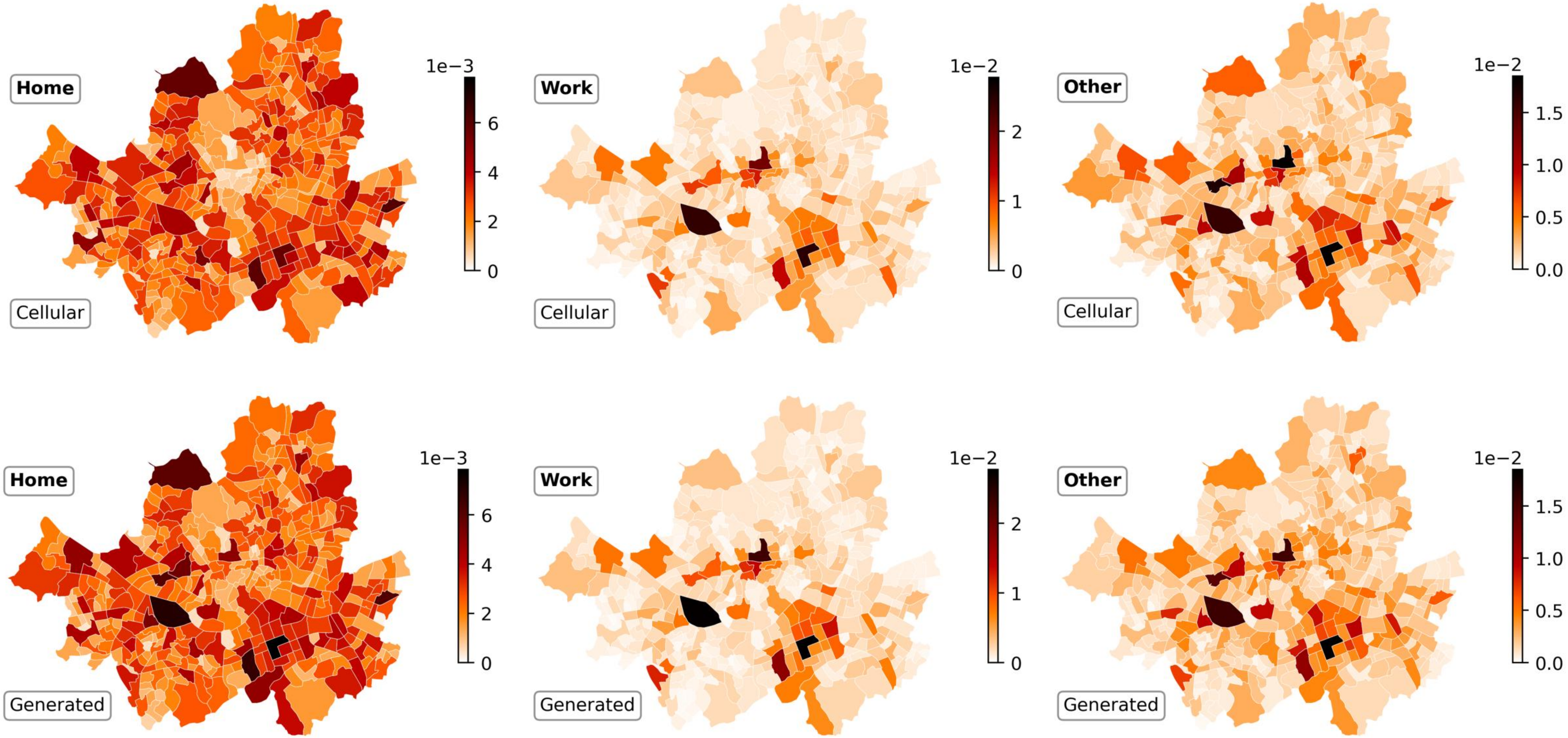
- **Spatiotemporal granularity**: The data fusion in Stage 1 can decide the efficient granularity of the feature space to bridge the PCM data with the HTS data.

Part 2: Validation (Singapore case study)



Fit of joint distributions between age and spatiotemporal attributes—arrival time, origin, and destination

Part 2: Validation (Seoul case study)



FUTURE CITIES LABORATORY GLOBAL

Behavioural COGNITIVE SCIENCE LAB

National University of Singapore

Questions?

LinkedIn Page

(Please follow for updates)