A new approach to generate daily activity schedules by fusing travel survey with time-dependent origin-destination matrices

Khoa D. Vo[ab], Prateek Bansal[ac*]
[a]Singapore-ETH Centre, Future Cities Lab Global Programme, Singapore Hub, Singapore
[b]Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam
[c]Department of Civil and Environmental Engineering, National University of Singapore, Singapore

Introduction

Emerging passively collected mobility (PCM) datasets–such as global positioning system traces, call detail records, and transit smart card transactions–offer an unprecedented opportunity to observe and understand mobility behavior in remarkable detail. These datasets have revolutionized our ability to reconstruct individual daily mobility patterns, but they have also sparked growing concerns over data privacy. In many cases, the use of such detailed data is restricted, necessitating the use of aggregated representations of travel demand. One widely used format for this purpose is the time-dependent origin-destination (TD-OD) matrix, which summarizes mobility as the total volume of trips between pairs of locations at different times of the day. Thanks to their simple structure, TD-OD matrices have become a cornerstone of transport planning and policymaking. However, their coarse-grained nature presents significant limitations. They fail to capture complex travel behaviors such as trip-chaining or interdependencies between trips. Moreover, the absence of sociodemographic details (e.g., age, gender) and trip information (e.g., trip-chain length, trip purposes) severely limits their application in agent-based models, which require richer data to simulate individual decision-making processes effectively.

To address the lack of sociodemographic and trip information in PCM data, we utilize household travel survey (HTS) data as a complementary source. Specifically, we propose a novel two-stage data fusion approach that leverages the TD-OD matrix from PCM data to generate comprehensive activity schedules for agents, while transferring trip information and sociodemographics from HTS data. In the first stage, we employ a cluster-based method (Vo et al., 2025) to generate a fused distribution that integrates the TD-OD matrix with sociodemographics and trip information. The second stage uses the fused distribution from the first stage to generate sociodemographics, trip information, and the spatiotemporal characteristics of activity schedules. This process leverages modified Markov models to improve the generation of activity schedules.

A significant limitation of conventional approaches (e.g., Anda et al., 2021; Ballis and Dimitriou, 2020; Ye et al., 2024) is their primary focus on the spatiotemporal aspects of activity schedules, often neglecting essential trip information such as trip chain length and trip purposes. This shortcoming frequently leads to the generation of numerous infeasible trip attribute combinations. Our proposed approach overcomes this limitation by integrating these critical trip attributes while preserving the distributions of sociodemographics and trip information from HTS data, as well as the spatiotemporal characteristics (i.e., TD-OD demands) from PCM data, through an effective data fusion process.

Methodology

Problem statement

Every agent, defined by their sociodemographic attributes X (e.g., age, gender), follows a daily activity schedule characterized by a sequence of activity purposes $Y = (\dots, Y^k, \dots)$, activity locations $Z = (\dots, Z_k, \dots)$, start times $S = (\dots, S_k, \dots)$, and end times $E = (\dots, E_k, \dots)$, where k denotes the order of trips in the activity chain. For individual trips, let $\bar{Z}, \tilde{Z}, \bar{T}$, and $\tilde{T}$ represent the origin, destination, departure time, and arrival time, respectively. The task is to harmonize two datasets: The HTS data, capturing the joint distribution $P(X = x, Y = y, Z = z, S = s, E = e)_{\text{hts}}$, which provides rich details about sociodemographics and trip information but suffers from low spatiotemporal heterogeneity (Vo et al., 2025). The PCM data in the TD-OD matrix form, describing $P(\bar{Z} = \bar{z}, \tilde{Z} = \tilde{z}, \bar{T} = \bar{t}, \tilde{T} = \tilde{t})_{\text{pcm}}$, which is highly reliable for spatiotemporal patterns but lacks critical contextual information. The data fusion problem aims to

estimate a unified joint distribution $P(X = x, Y = y, Z = z, S = s, E = e)_{\text{fus}}$, combining the strengths of both datasets. For simplicity, we write these distributions as $p(x, y, z, s, e)_{\text{hts}}$ for HTS data and $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{pcm}}$ for PCM data.

Preliminary data fusion approach

A conventional method for addressing the data fusion problem involves a preliminary two-stage approach, which has been widely adopted in various forms in existing literature (e.g., Anda et al., 2021; Ballis and Dimitriou, 2020; Ye et al., 2024). The first step in solving the data fusion problem is to harmonize the spatiotemporal information between the datasets by generating $p(z, s, e)_{\text{pcm}}$ from $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{pcm}}$. To streamline the mathematical representation, we can express $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ from PCM data in terms of trip chain order to make it consistent with $p(z_{k-1}, e_{k-1}, z_k, s_k)$ in HTS data. This generation process can be effectively modeled using the following Markov process:

$$p(z, s, e)_{\text{pcm}} = p(z_1, s_1)_{\text{pcm}} p(e_1|z_1, s_1)_{\text{pcm}} \prod_{k=2}^{K} p(z_k, s_k|z_{k-1}, e_{k-1})_{\text{pcm}} p(e_k|z_k, s_k)_{\text{pcm}}. \tag{1}$$

The second step involves the transfer of spatiotemporal information (z, s, e) between the datasets and enables the imputation of (x, y) from HTS data into (z, s, e) in PCM data. This process is expressed as:

$$p(x, y, z, s, e)_{\text{fus}} = p(x, y|c)_{\text{fus}} p(z, s, e, c)_{\text{fus}} = \frac{p(x, y, c)_{\text{hts}}}{p(c)_{\text{fus}}} p(z, s, e, c)_{\text{pcm}} \tag{2}$$

where c represents a spatiotemporal cluster that maps (z, s, e) across both datasets. In Eq. (2), $p(x, y, c)_{\text{hts}}$ and $p(z, s, e, c)_{\text{pcm}}$ are derived from HTS and PCM data, respectively, capturing the strengths of each dataset.

The preliminary two-stage approach has the following challenges:

Feasibility: Existing methods (e.g., Anda et al., 2021; Ballis and Dimitriou, 2020; Ye et al., 2024) that adopt the generation process described in Eq. (1) often produce a significant number of infeasible combinations (z, s, e). This issue arises due to the sequential nature of the process and the lack of critical trip-chain information–such as trip chain length and trip purposes–when generating the next activity location and arrival time $(z_k, s_k|z_{k-1}, e_{k-1})$, as well as the end time at the next location $(e_k|z_k, s_k)$.

Distribution preservation: The data fusion process described in Eq. (2) relies on different assumptions about the distributions preserved in the fused output. Specifically, setting $p(c)_{\text{fus}} = p(c)_{\text{hts}}$ ensures that $p(z, s, e)_{\text{fus}} = p(z, s, e)_{\text{pcm}}$, which also preserves $p(z_{k-1}, e_{k-1}, z_k, s_k)_{\text{fus}} = p(z_{k-1}, e_{k-1}, z_k, s_k)_{\text{pcm}}$ at an aggregate level. Conversely, setting $p(c)_{\text{fus}} = p(c)_{\text{pcm}}$ results in $p(x, y)_{\text{fus}} = p(x, y)_{\text{hts}}$. However, ideally, the fused distribution should *simultaneously* maintain, to some extent, both $p(x, y)_{\text{fus}} = p(x, y)_{\text{hts}}$ and $p(z_{k-1}, e_{k-1}, z_k, s_k)_{\text{fus}} = p(z_{k-1}, e_{k-1}, z_k, s_k)_{\text{pcm}}$.

Spatiotemporal granularity: The data fusion process described in Eq. (2) relies on assumptions about the granularity of clusters c to bridge (z, s, e) from the PCM data with (x, y) from the HTS data. However, the number of possible combinations of (z, s, e) in the PCM data is significantly higher than in the HTS data, making this alignment challenging. Deep generative models offer a potential solution by learning a latent space that functions similarly to clusters c, enabling the connection between (z, s, e) in the PCM data and (x, y) in the HTS data. However, these models fail to ensure the preservation of distributions, which is critical for maintaining the integrity of both datasets in the fused result.

Proposed data fusion approach

We modify the activity generation process in Eq. (1) as follows:

$$p(x, y, z, s, e)_{\text{fus}} = p(x, y, z_1, s_1)_{\text{fus}} p(e_1|x, y, z_1, s_1)_{\text{fus}} \prod_{k=2}^{K} p(z_k, s_k|x, y, z_{k-1}, e_{k-1})_{\text{fus}} p(e_k|x, y, z_k, s_k)_{\text{fus}}. \tag{3}$$

The advantage of the generation process in Eq. (3) over that in Eq. (1) lies in its incorporation of sociodemographics and trip-chain information when determining the next activity location and start time $(z_k, s_k|x, y, z_{k-1}, e_{k-1})$ as well as the end time $(e_k|x, y, z_k, s_k)$. By utilizing this additional contextual information, the process addresses the feasibility challenge by significantly reducing the number of possible spatiotemporal combinations for the next activity location, along with its associated start and end times.

However, the generation process in Eq. (1) depends on $p(x, y, z_{k-1}, e_{k-1}, z_k, s_k)_{\text{fus}}$. Since $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ and $(z_{k-1}, e_{k-1}, z_k, s_k)$ in Eq. (3) are interchangeable, we reformulate the data fusion problem as finding $p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}$ such that:

$$\min \sum_{x\in\Gamma(X)} \sum_{y\in\Gamma(Y)} \sum_{c\in\Gamma(C)} \mathcal{D}_{\text{JS}} \left(p(x, y, c)_{\text{hts}} \| p(x, y, c)_{\text{fus}}\right) \tag{4}$$

subject to

$$p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} = p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{pcm}} \tag{5}$$

$$p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} \geq 0 \tag{6}$$

where c is now a cluster of $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ at the trip level instead of (z, s, e) at the activity-schedule level, and $\Gamma(.)$ is the set of possible value combinations for the given attributes. Note that we have $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}} = p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}$ as $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ encompasses all information about c.

We exclude $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ from the objective function in Eq. (4) due to the low spatiotemporal heterogeneity of the HTS data. Instead, a clustering-based approximation that leverages c in place of $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ is necessary to align both datasets to a consistent spatiotemporal granularity for effective information transfer. The objective function in Eq. (4) minimizes the Jensen-Shannon (JS) divergence between the fused joint distribution of (x, y, c) and that of the HTS data. Constraint (5) indicates that the fused distribution of $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ is the same as of the PCM data. Constraint (6) ensures that the probability mass on each attribute combination is non-negative.

The data fusion process outlined in Eqs. (4)–(6) addresses the challenge of distribution preservation, ensuring the retention of $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{pcm}}$ and $p(x, y)_{\text{hts}}$. However, it still faces the spatiotemporal-granularity challenge, particularly regarding the assumptions made when creating clusters. In addition, the data fusion process introduces a new challenge related to *high dimensionality*—the combinatorial explosion of possible combinations for $(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)$ across the HTS and PCM datasets. Solving the optimization problem directly is computationally infeasible due to the sheer number of decision variables, which scales as $O(|\Gamma(X)|\times|\Gamma(Y)|\times|\Gamma(\bar{Z})|\times|\Gamma(\tilde{Z})|\times|\Gamma(\bar{T})|\times|\Gamma(\tilde{T})|)$, where $|\Gamma(\cdot)|$ denotes the size of the corresponding attribute's domain. To address this issue, we propose an approximate but tractable reformulation to estimate $p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}$. This reformulation is equivalent to the original problem defined in Eqs. (4)–(6), ensuring that the high-dimensional optimization problem is reduced to multiple low-dimensional subproblems. These subproblems can be efficiently solved, making the approach computationally feasible while preserving the fidelity of the original optimization goals.

Reformulated equivalent data fusion problem

We now reformulate the optimization problem described in Eqs. (4)–(6) to deal with the high-dimensionality challenge. We can express the fused joint distribution $p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}}$ as:

$$p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}} = p(x, y|\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}} p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}}. \tag{7}$$

We also have $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{fus}} = p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}$ as $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ includes all information about c. Thus, Eq. (7) can be rewritten:

$$p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} = p(x, y|\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}. \tag{8}$$

Similar to the original problem given in Eqs. (4)–(6), we need to rely on c instead of $(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})$ in the conditional distribution to ensure the same granularity of both datasets for information transfer:

$$p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} \approx p(x, y|c)_{\text{fus}} p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} = \frac{p(x, y, c)_{\text{fus}}}{p(c)_{\text{fus}}} p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}}. \tag{9}$$

We also hypothesize that $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{pcm}}$ is more reliable than $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t})_{\text{hts}}$. Thus, $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{pcm}}$ must be also more reliable. We rewrite the joint distribution in Eq. (9) as

$$p(x, y, \bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} \approx \frac{p(x, y, c)_{\text{fus}}}{p(c)_{\text{pcm}}} p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{pcm}}. \tag{10}$$

Similar to the objective function in Eq. (4), we would like to minimize the probabilistic distance between $p(x, y, c)_{\text{fus}}$ and $p(x, y, c)_{\text{hts}}$ but we cannot impose $p(x, y, c)_{\text{fus}} = p(x, y, c)_{\text{hts}}$ in Eq. (10). Directly replacing $p(x, y, c)_{\text{fus}}$ with $p(x, y, c)_{\text{hts}}$ can distort $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{pcm}}$ and causes $p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{fus}} \neq p(\bar{z}, \tilde{z}, \bar{t}, \tilde{t}, c)_{\text{pcm}}$. To address this issue, $p(x, y, c)_{\text{fus}}$ can be obtained as the solution to the following optimization problem $P_{\text{opt\_fusion}}$:

$$\min \sum_{x \in \Gamma(X)} \sum_{y \in \Gamma(Y)} \sum_{c \in \Gamma(C)} \mathcal{D}_{\text{JS}} \left(p(x, y, c)_{\text{fus}} || p(x, y, c)_{\text{hts}} \right) \tag{11}$$

subject to

$$\sum_{x \in \Gamma(X)} \sum_{y \in \Gamma(Y)} p(x, y, c)_{\text{fus}} = p(c)_{\text{pcm}} \tag{12}$$

$$p(x, y, c)_{\text{fus}} \geq 0. \tag{13}$$

The objective function in Eq. (11) is the JS divergence between $p(x, y, c)_{\text{fus}}$ and that from the HTS data conditional on $p(c)_{\text{pcm}}$. The conditional constraint (12) preserves the distribution of c from the PCM data. The reformulated problem can be decomposed into $O(|\Gamma(C)|)$ subproblems—one for each cluster, each of which has $O(|\Gamma(X)| \times |\Gamma(Y)|)$ decision variables. For more details on the properties of the data fusion problem and strategies for creating effective clusters, readers are referred to Vo et al., 2025.

What's next?

We will conduct two experiments to evaluate the proposed framework: (i) validation of its effectiveness, and (ii) demonstration of its practical application through a real-world case study. For these experiments, we will utilize a TD-OD matrix synthesized from HTS data and various PCM data sources, including smart card and taxi data collected in Singapore in July 2022. In the validation experiment, we use the "full HTS data" as the ground truth. From this dataset, we sample 5% of the observations to create the "hypothetical HTS data" and use all trips to create TD-OD matrices obtained from "hypothetical PCM data." The purpose of this validation step is to assess how effectively the proposed framework can integrate the hypothetical HTS data and the hypothetical TD-OD matrix to reconstruct activity schedules and sociodemographics that align with the full HTS data (i.e., the ground truth). In the case study, we will apply the framework to actual HTS data and a real TD-OD matrix for Singapore. External dataset will be used for validation.

References

Anda, C., Medina, S. A. O., & Axhausen, K. W. (2021). Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. *Transportation Research Part C: Emerging Technologies*, *128*, 103118.

Ballis, H., & Dimitriou, L. (2020). Revealing personal activities schedules from synthesizing multi-period origin-destination matrices. *Transportation research part B: Methodological*, *139*, 224–258.

Vo, K. D., Kim, E. J., & Bansal, P. (2025). A novel data fusion method to leverage passively-collected mobility data in generating spatially-heterogeneous synthetic population. *Transportation Research Part B: Methodological*, *191*.

Ye, J., Hu, Y., & Gao, L. (2024). Data-driven framework for generating travelers with demographic-activity-travel information. *Transportation Research Record: Journal of the Transportation Research Board*, *2678*, 1–13.