NATIONAL CENTER FOR UNDERSTANDING FUTURE
TRAVEL BEHAVIOR AND DEMAND

Final Project Report

# Imputing Socio-Demographics for Mobile Trajectories

***BY***

**Cynthia Chen**
Email: qzchen@uw.edu

**Ekin Ugurel**
Email: ugurel@uw.edu

Department of Civil and Environmental Engineering
University of Washington
1410 NE Campus Pkwy, Seattle, WA 98195

March 2026

## TECHNICAL REPORT DOCUMENTATION PAGE

| | | | |
|---|---|---|---|
| **1. Report No.**<br>N/A | **2. Government Accession No.**<br>N/A | **3. Recipient's Catalog No.**<br>N/A | |
| **4. Title and Subtitle**<br>Imputing Socio-Demographics for Mobile Trajectories | | **5. Report Date**<br>March 11, 2026 | |
| | | **6. Performing Organization Code**<br>N/A | |
| **7. Author(s)**<br>Cynthia Chen, Ph.D., https://orcid.org/0000-0002-1110-8610<br>Ekin Ugurel, https://orcid.org/0000-0002-0849-4688 | | **8. Performing Organization Report No.**<br>N/A | |
| **9. Performing Organization Name and Address**<br>Department of Civil and Environmental Engineering<br>University of Washington<br>1410 NE Campus Pkwy, Seattle, WA 98195 | | **10. Work Unit No. (TRAIS)**<br>N/A | |
| | | **11. Contract or Grant No.**<br>69A3552344815 and 69A3552348320 | |
| **12. Sponsoring Agency Name and Address**<br>U.S. Department of Transportation,<br>University Transportation Centers Program,<br>1200 New Jersey Ave, SE, Washington, DC 20590 | | **13. Type of Report and Period Covered**<br>Final Report, 2024-2025 | |
| | | **14. Sponsoring Agency Code**<br>USDOT OST-R | |
| **15. Supplementary Notes**<br>N/A | | | |


**16. Abstract**

Inferring sociodemographic attributes from mobility data could help transportation planners better leverage passively collected datasets, but this task remains difficult due to weak and inconsistent relationships between mobility patterns and sociodemographic traits, as well as limited generalization across contexts. We address these challenges from three angles. First, to improve predictive accuracy while retaining interpretability, we introduce a behaviorally grounded set of higher-order mobility descriptors based on directed mobility graphs. These features capture structured patterns in trip sequences, travel modes, and social co-travel, and significantly improve prediction of age, gender, income, and household structure over baselines features. Second, we introduce metrics and visual diagnostic tools that encourage evenness between model confidence and accuracy, enabling planners to quantify uncertainty. Third, to improve generalization and sample efficiency, we develop a multitask learning framework that jointly predicts multiple sociodemographic attributes from a shared representation. This approach outperforms single-task models, particularly when training data are limited or when applying models across different time periods (i.e., when the test set distribution differs from the training set).


| | | | |
|---|---|---|---|
| **17. Key Words**<br>Demographic Inference; Household Travel Survey; Multi-task Learning; Uncertainty Quantification | | **18. Distribution Statement**<br>No restrictions. | |
| **19. Security Classif.(of this report)**<br>Unclassified | **20. Security Classif.(of this page)**<br>Unclassified | **21. No. of Pages**<br>46 | **22. Price**<br>N/A |

## DISCLAIMER

*The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under Grant No. 69A3552344815 and 69A3552348320 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.*


## ACKNOWLEDGMENTS

This research was partially supported by the National Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the U.S. Department of Transportation (USDOT) under grant numbers 69A3552344815 and 69A3552348320. The authors would like to thank the TBD National Center and USDOT for their support of university-based research in transportation, particularly for the funding provided for this project. The authors also extend their thanks to the Puget Sound Regional Council for their valuable contributions to the work presented in this report, including the three waves of the household travel survey which was the main dataset used in our experiments.

**TABLE OF CONTENTS**

## LIST OF TABLES



## LIST OF FIGURES

## EXECUTIVE SUMMARY

The rapid proliferation of GPS-enabled smartphones and location-based services has provided transportation planners with massive, continuously refreshed datasets that capture where and when people move. However, unlike traditional Household Travel Surveys (HTS), these passive data sources almost exclusively lack sociodemographic information—the "who" behind the movement. This missing dimension severely limits the ability of Metropolitan Planning Organizations (MPOs) to conduct distributional analyses, assess equity impacts, or model travel behavior differences across population subgroups. While previous research has attempted to infer these attributes from trajectory data, results have often been hindered by weak correlations between simple mobility metrics and sociodemographics, as well as a lack of model transferability across different geographic or temporal contexts. This report addresses these challenges by developing a comprehensive framework for inferring sociodemographic attributes (age, gender, income, and household structure) using behaviorally grounded mobility descriptors and multitask deep learning.

To improve the predictive accuracy and interpretability of inference models, we introduce a family of "higher-order" mobility descriptors derived from directed mobility graphs. In this framework, vertices represent activity purposes and edges represent trip sequences. Beyond simple frequency counts, we extract complex features such as "motifs" (recurring daily subgraph patterns like chains or loops), global and local clustering coefficients (measures of network cohesion), and social co-travel composition (solo vs. group travel). Our analysis of the Puget Sound Regional Council (PSRC) Household Travel Survey data reveals that these descriptors carry significant sociodemographic signals; for example, complex tour structures and specific co-travel patterns are strongly correlated with household size and life-cycle stages. When integrated into predictive models, these higher-order features consistently improve out-of-sample accuracy and likelihood scores compared to baseline spatiotemporal features.

A critical innovation of this study is the application of Multitask Learning (MTL) to enhance model generalization and data efficiency. Rather than training separate models for each demographic attribute, we utilize a shared-trunk neural network architecture that jointly predicts age, gender, income, and number of children. By sharing a latent representation of mobility behavior across related tasks, the model effectively pools statistical evidence, which is particularly beneficial when training data is scarce or when the model is applied to data that differs distributionally from the training set. We validated this approach using a "cross-temporal" experimental design, training on pre-pandemic survey waves (2017, 2019) and testing on post-pandemic data (2021, 2023). The results demonstrate that the multitask framework consistently outperforms single-task baselines in data-sparse regimes, exhibiting greater robustness to the behavioral shifts caused by the COVID-19 pandemic.

Furthermore, this report emphasizes the importance of uncertainty quantification in demographic inference. Standard accuracy metrics often obscure the reliability of predictions, leading to potential overconfidence in decision-making. We operationalize calibration metrics, specifically Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL), alongside visual diagnostic tools like reliability diagrams to ensure that model confidence aligns with observed accuracy. Our experiments indicate that while adding complex features improves separability (AUROC), it does not always guarantee better calibration; however, the multitask learning framework aids in regularizing these probability estimates, preventing overconfidence even under

distribution shifts.

In conclusion, this research provides a robust methodological pathway for enriching passive mobility data with sociodemographic attributes. We demonstrate that combining behaviorally meaningful graph descriptors with multitask learning architectures significantly enhances both the accuracy and the reliability of demographic inference. While the study relies on processed survey diaries to establish ground truth, the features developed are compatible with raw GPS traces once activity purposes are imputed. These tools enable planners to leverage the scale of big data while retaining the sociodemographic context necessary for equitable and insightful transportation planning.

## 1. INTRODUCTION

Over several decades, travel behavior scholarship has revolved around the interplay between *who* a person is and *how* they move. Household-travel surveys (HTSs) have long provided the empirical backbone for this work by coupling rich trip diaries with respondent characteristics such as age, gender, income, and household composition. Analyses drawing on these surveys consistently show that, after accounting for the built environment, sociodemographic traits still correlate with car ownership, mode choice, trip frequency, and trip-chaining behavior (Bhat and Misra, 1999; Lee et al., 2007; Lu and Pas, 1999; McGuckin and Murakami, 1999; Mokhtarian and Chen, 2004)

In the past dozen years, the ubiquity of GPS-enabled smartphones has spawned a parallel, industry-scale source of mobility evidence in passively-generated mobile data, which includes call-detail records (CDR), location-based service (LBS) pings, connected-vehicle traces, and the like (Chen et al., 2016). These datasets dwarf HTSs in both sample size and temporal length, are refreshed continuously, and can often be licensed at a fraction of the cost of running a tailored survey. Their content, however, is almost exclusively spatial temporal; they record where and when a device was observed but remain agnostic about *who* was holding it. This missing dimension limits many distributional and behavioral analyses, including those that require understanding how travel patterns vary across population subgroups.

Despite this blind spot, public agencies have been keen on experimenting with mobile data products (Ugurel et al., 2024). Metropolitan planning organizations (MPOs) see potential in using them to stitch origin-destination matrices (Alexander et al., 2015; Iqbal et al., 2014), site electric-vehicle chargers (Yang et al., 2017), and evaluate complete-street retrofits (Bian et al., 2023). Yet the lack of respondent attributes imposes two related hazards. First, representativeness: smartphone datasets systematically under sample certain population subgroups (Li et al., 2024; Wang et al., 2025; Wesolowski et al., 2013; Wu et al., 2024b). As a result, decisions based solely on such data may reflect the travel patterns of overrepresented groups while ignoring others. Second, for many analyses MPOs would like to conduct, such as gauging the effects of new toll roads or assessing if expanded transit lines reach underserved communities, linking mobility traces to travelers' sociodemographic profiles is required.

To unlock the full value of passively-generated traces, researchers need a way to infer or impute sociodemographic variables from the mobility patterns embedded in the devices. From a modeling perspective, this represents an "inverse" problem. While travel behavior is traditionally modeled as a function of sociodemographics, $T = f(S)$, we seek to characterize the probabilistic mapping $P(S|T)$. We acknowledge that this mapping is rarely deterministic--multiple sociodemographic profiles can produce similar mobility traces. However, by leveraging higher-order behavioral signals, we can significantly narrow the range of likely sociodemographic profiles for a given traveler compared to simply using population averages.

This issue has been studied from a variety of angles, including travel behavior (Auld et al., 2015; Zhang et al., 2024), data mining of destination choice (Doi et al., 2021; Solomon et al., 2018; Wu et al., 2019; Zhong et al., 2015), communication metadata (Jahani et al., 2017; Razavi et al., 2024), transit smart card use (Ding et al., 2019), and information theory (Zhao et al., 2022). As our goal is to enable the use of mobile data for transportation planning, we primarily focus on studies that solely use mobility behavior to achieve the aim of imputing sociodemographic attributes from mobility traces.

Investigating this sociodemographic-inference problem presents two principal obstacles. First, most of the foundational literature in travel behavior frames sociodemographics as shaping mobility, not the reverse. This perspective aligns with behavioral models in which age, income, and household responsibilities influence when, where and how people travel. As a result, conceptualizing mobility traces as predictive signals for inferring sociodemographic attributes remains a relatively underdeveloped and counterintuitive direction in the literature. Second, sociodemographic inference from behavioral signals is intrinsically difficult. While some features of travel (e.g., trip chaining or mode diversity) can correlate with variables like gender or income, the relationships are often weak, noisy, and highly context dependent. In practice, models trained on one dataset may exhibit strong predictive performance yet fail to generalize when applied to another dataset, especially across geographies with different urban forms and social norms (Sheller and Urry, 2006). These challenges complicate efforts to construct transferable models and limit the extent to which empirical findings can be applied universally.

This study aims to overcome these obstacles through the following contributions. First, to confront the limited theoretical grounding for inferring demographics from mobility, we introduce a family of higher-order descriptors based on directed mobility graphs in which vertices encode activity purposes and edges connect chronologically adjacent trips. By *higher*-order, we mean measures that go beyond first-order metrics (i.e., visit counts, mode frequencies) to capture relations across sequences of trips and destinations (e.g., how evenly travel is spread across activities, whether trips form loops or tours, whether they mix modes, etc.). We demonstrate that these descriptors are strongly and interpretably associated with sociodemographic attributes and that they substantially increase the predictive power of imputation models beyond classical and spatiotemporal features (defined in Section 4.1).

Second, to mitigate the inherent difficulty and uncertainty in generalizing sociodemographic inference models across diverse contexts, we operationalize uncertainty quantification and model calibration for multi-class classification. Specifically, we leverage metrics that encourage models to match their prediction confidences with their accuracies. We also use visual diagnostic tools like reliability diagrams to identify calibration gaps in our experiments. Applying this protocol, we find that the proposed higher-order descriptors consistently improve out-of-sample likelihoods, but their effect on calibration is mixed: they close gaps in several settings but can yield conservative (under-confident) probabilities, possibly due to the risk of overfitting.

Third, we establish the value of a multitask (MT) learning strategy that predicts multiple sociodemographic attributes simultaneously. Under standard statistical learning assumptions, parameter sharing across related tasks reduces estimator variance and improves out of sample performance (Baxter, 2000; Caruana, 1997). Thus, jointly modeling targets like age, gender, income, and household size allows the network to exploit shared structure in the mapping from mobility to sociodemographic attributes, yielding a more data-efficient and robust estimator than training separate models. Moreover, MT learning can improve transferability across contexts by reducing sensitivity to distributional shift (i.e., changes in the input–output relationship between training and test data). We corroborate these claims with cross-temporal generalization experiments in which the multi-task variant consistently outperforms single-task baselines under a range of sample-size constraints.

**Our Contributions:**

- We introduce a behaviorally grounded family of **higher-order descriptors based on mobility graphs**, in which vertices represent travel purposes and edges represent temporally ordered trips between them. When appended to classical mobility features, our feature set consistently raises the out-of-sample accuracy of sociodemographic inference (e.g., age, gender, income, etc.) across multiple experimental setups. We further evaluate their robustness through a sensitivity analysis that disentangles the effects of attribute coarsening and inference bias arising from GPS-based enrichment, providing empirical evidence that the proposed descriptors retain predictive value under realistic passive-data conditions.
- We operationalize **uncertainty quantification and calibration methods** for mobility-based sociodemographic inference, using metrics that align predicted confidence with observed accuracy. Visual tools like reliability diagrams reveal calibration gaps and help diagnose model behavior. We discuss practical use cases in which well-calibrated probabilities are essential, including propagating imputation uncertainty into downstream behavioral models or reporting accessibility metrics with uncertainty bounds rather than deterministic group assignments.
- We demonstrate the benefits of **multi-task learning** for transferability and data efficiency. Training a unified model to predict multiple sociodemographic attributes jointly improves sample efficiency and enhances generalization to new data compared to single-task baselines. In both data sparse regimes and those in which the test set systematically differs from the training set, the multi-task variant consistently outperforms single-task counterparts, indicating enhanced robustness and practical transfer across contexts.

The rest of this report is organized as follows: Section 2 provides a review of relevant literature, both in the context of classical travel behavior studies and those that seek to answer the same question as ours. Section 3 outlines the datasets we leverage, while Section 4 details our methodological approach. Section 5 presents our experimental design and the numerical results. We conclude with a discussion of our findings, limitations, and ideas for future work in Section 6.

## 2. LITERATURE REVIEW

In this section, we review previous efforts to infer sociodemographic backgrounds of individuals based on their mobility behavior. We distinguish between these studies and a parallel body of work

that predicts demographics from mobile phone communication metadata (call records, texting patterns, app usage, etc.). Both lines of research share the premise that digital behavioral traces contain sociodemographic signals, but they differ in the type of behavior analyzed. Mobility focused studies use *where people go* as the primary predictor, whereas communication focused studies use who people connect with and how they use their devices. Beyond individual-level inference, a related branch of scholarship focuses on data fusion and population synthesis to enrich mobility data. These approaches often leverage aggregate small-area statistics from the census or samples of anonymized records to estimate population microdata (Ryan et al., 2009; Williamson et al., 1998). More recent work has utilized passively-collected mobility data to generate spatially-heterogeneous synthetic populations, validating individual-level behavioral inferences against ground-truth census distributions (Vo et al., 2025). While our work focuses on direct behavioral inference, the descriptors and multitasking framework we propose could potentially enhance these fusion models by providing more nuanced behavioral constraints. Census data and other sources providing "ground-truth" statistics can also be used to narrow the conditional posterior distribution of sociodemographic attributes in a Bayesian sense, effectively reducing predictive uncertainty by grounding individual-level mobility signals within known population-level statistics.

Most mobility-inference studies begin by transforming raw coordinates into interpretable spatial or temporal descriptors. Spatial metrics quantify the extent and diversity of an individual's movements: the radius of gyration (Ding et al., 2019), the heterogeneity of visited locations (Wu et al., 2019), the number of unique visited locations (Wu et al., 2019; Zhong et al., 2015), and those that relate to the distance traversed (Wu et al., 2019). Temporal features characterize regularity and rhythm, such as the day-to-day similarity of location sequences or the distribution of departure times for commutes (Ding et al., 2019), under the hypothesis that highly routinized commuters differ demographically from, for example, students or gig-economy workers. Semantic signals add further discriminatory power. Studies extract the frequency of visits to certain points-of-interest (POIs). For example, frequent stops at beauty salons or supermarkets can proxy for gender (Doi et al., 2021), while regular, short school drop-offs can signal parental status. From a classical travel behavior lens, stop durations, tour counts, and land-use contexts also prove to carry predictive power on whether or not a person drives, their work status, and education level (Auld et al., 2015). Despite the range of mobility descriptors, it remains unclear which signals reliably predict which attributes and under what conditions. Reported associations are often ad hoc and context-specific. Most studies focus on overall accuracy using bundled features, without isolating the marginal value of specific groups (spatial, temporal, semantic) or testing their consistency across settings. To fill this gap, we systematically assess the incremental contribution and robustness of each feature family.

To interpret these descriptors, researchers have moved from statistical models to more data-driven approaches. Early work mapped those hand-crafted features into conventional statistical or rule-based models. Auld et al. (2015) combined fuzzy clustering of classical travel descriptors with decision trees and nested-logit models, while Zhong et al. (2015) used tensor factorization to decompose location check-in patterns. More recent research has gravitated toward representation-learning. Solomon et al. (2018) interpreted each day's trajectory as a "sentence'' and learned Word2Vec embeddings prior to classification, whereas Ding et al. (2019) employed long short-term memory networks on smart-card sequences. Xu et al. (2020) modeled the city as a heterogeneous mobility network and learned embeddings by preserving both physical co-visitation and semantic similarity between users. Generally, these approaches aim to capture the temporal

structure and contextual regularities of daily movement, letting the model implicitly learn what aspects of movement are informative for demographics. However, most of these models train separately for each attribute, ignoring shared structure across related targets (i.e., age and the number of children). This makes them data-hungry and less transferable across settings with distribution shifts. We address this by using a multitask learning framework that jointly predicts all attributes from a shared representation, improving sample efficiency and out-of-sample robustness.

When it comes to the output of prediction models, the uncertainty that is quantified needs to be interpreted with care. Many works simply present a confusion matrix or error rate, which is an aggregate uncertainty measure. These tell us, for example, that the model is wrong 20% of the time overall, but not *which* 20\% of cases or how to flag an uncertain individual. The tendency of earlier models to overfit their training data (as in Auld et al., 2015) highlights the risks of relying on point predictions without accounting for variance or confidence. A few studies have taken steps toward more meaningful uncertainty estimates. For example, some inference models incorporate cross-validation accuracy into probability estimates, essentially adjusting predicted class probabilities downward to reflect the model's known error rate (Zhang et al., 2024). This can prevent overconfidence in the predictions by "baking in" the chance of error. Moreover, the recent work by Zhao et al. (2022) introduces a theoretical framework to estimate beforehand how separable the classes might be, given the covariance structure of mobility data. Building on these advances, we calibrate predicted probabilities using metrics that encourage honesty between confidence and accuracy. We further show that our feature set and multitask learning framework yield more reliable and discriminative uncertainty estimates compared to baselines. These steps support a shift toward uncertainty-aware inference, where decisions can reflect the model's confidence.

## 3. DATASET

We analyze three of the four most recent waves of the Puget Sound Regional Council (PSRC) Household Travel Survey (HTS), with the exception being the 2021 wave during which travel behavior was heavily confounded by the COVID-19 pandemic. The 2017, 2019, and 2023 surveys were fielded biennially using an address-based probability sample covering the four-county central

Puget Sound region (King, Kitsap, Pierce, and Snohomish). The 2021 wave uniquely included an additional opt-in online panel, which we omit here for consistency. Table 1 highlights relevant descriptive statistics from our post-processed version of this dataset.

**Table 1 Descriptive statistics of the PSRC HTS in the three waves used in this study (post-processing; values in parentheses denote the strata percentage associated with wave)**

| | **Wave** | **2017** | **2019** | **2023** |
|---|---|---|---|---|
| | Field Dates | 04/10 – 06/15 | 03/11 – 05/30 | 04/24 – 05/29 |
| | Households | 3,160 | 2,902 | 3,504 |
| | Persons | 5,545 | 5,116 | 5,959 |
| | Trips | 51,029 | 71,913 | 56,028 |
| Gender | Male | 2,714 (**48.9**%) | 2,475 (**48.4**%) | 2,597 (**43.6**%) |
| | Female | 2,724 (**49.1**%) | 2,533 (**49.5**%) | 2,854 (**47.9**%) |
| | Non-binary | 17 (**0.31**%) | 23 (**0.45**%) | 121 (**2.03**%) |
| Age | 0-11 | 495 (**9.74**%) | 477 (**10.1**%) | 549 (**10.8**%) |
| | 12-17 | 151 (**2.97**%) | 159 (**3.36**%) | 197 (**3.93**%) |
| | 18-34 | 1,739 (**34.2**%) | 1,514 (**31.9**%) | 1,335 (**26.7**%) |
| | 35-54 | 1,562 (**30.7**%) | 1,480 (**31.3**%) | 1,517 (**30.3**%) |
| | 55-74 | 970 (**19.1**%) | 941 (**19.9**%) | 1,152 (**23.0**%) |
| | 75+ | 165 (**3.25**%) | 163 (**3.44**%) | 267 (**5.33**%) |
| Household Income | Under $25,000 | 410 (**8.07**%) | 314 (**6.63**%) | 354 (**7.07**%) |
| | $25,000-$49,999 | 642 (**12.6**%) | 588 (**12.4**%) | 533 (**10.6**%) |
| | $50,000-$74,999 | 744 (**14.6**%) | 715 (**15.1**%) | 646 (**12.9**%) |
| | $75,000-$99,999 | 728 (**14.3**%) | 657 (**13.9**%) | 531 (**10.6**%) |
| | $100,000+ | 2,558 (**50.3**%) | 2,460 (**51.9**%) | 2,943 (**58.8**%) |
| Number of Children in Household | 0 | 3,538 (**69.6**%) | 3,297 (**69.6**%) | 3,371 (**67.3**%) |
| | 1 | 702 (**13.8**%) | 586 (**12.3**%) | 604 (**12.1**%) |
| | 2 | 712 (**14.0**%) | 601 (**12.7**%) | 755 (**15.1**%) |
| | 3+ | 130 (**2.57**%) | 250 (**5.28**%) | 277 (**5.53**%) |

During data cleaning we applied a set of exclusion rules to ensure that only complete and internally consistent trip records entered the analysis file. First, any trip lacking a valid origin or destination purpose code was deleted, because purpose is central to several of our behavioral indicators. We likewise removed trips for which the travel-mode field was blank; mode choice is a key outcome variable and cannot be reliably imputed when entirely missing. Third, observations without usable spatial information—specifically, trips whose destination coordinate could not be geocoded or whose reported distance was zero—were discarded, as they preclude computation of distance-based measures. Finally, we excluded the small number of records reporting negative travel durations, which are symptomatic of data-entry or time-stamp errors. These filters leave a sample of trips with complete purpose, mode, spatial and temporal attributes suitable for subsequent modelling.

Although our motivation is to enable demographic inference from passively-collected mobile data, we conduct this study using HTS data due to two key advantages. First, it provides ground-truth sociodemographic labels, which are essential for supervised learning and evaluation. Second, because HTS is fielded biannually, it enables testing under distribution shift by evaluating models across different survey waves. While our features are derived from structured trip diaries, recent advances in imputation algorithms now make it feasible to extract similar semantic information from raw GPS traces (Gao et al., 2024; Merikhipour et al., 2024). Thus, the methods developed here can largely be applied to passive data once suitably enriched. In Section 5.4, we approximate trip purposes in PSRC's GPS-based subset via reverse

geocoding and land-use/POI context around activity locations, while also estimating travel modes from speed and acceleration profiles along trajectories. These inferred labels are inherently noisier and typically defined at a coarser semantic resolution than their HTS counterparts.

## 4. METHODOLOGY

Section 4 details our methodology. 4.1 formalizes the mobility-graph representation and defines the activity- and trip-level covariates used in prediction. In 4.2, we outline definitions and metrics that allow us to operationalize uncertainty quantification in our context. Finally, 4.3 describes the theory behind the MT approach as well as the specific neural architecture we leverage.

### 4.1 Characterizing travel behavior with mobility graphs

Daily activity chains are encoded as directed graphs $G = (V, E)$ in which vertices $V$ represent unique destination purposes (e.g., home, gym, school), and edges $E$ represent temporally ordered trips between them. From this representation we extract two layers of information: (1) **activity (node) features**, referring to how frequently, evenly, and in which sequence specific activities are pursued, and (2) **trip (edge) features**, which summarize mode diversity (tendency to use different travel modes), co-travel composition (share of trips taken alone vs. with others), and daily travel motifs (minimal, recurring subgraphs that capture the daily structure of trip sequences). We give precise definitions for metrics to quantify the above in Section 4.1.2 (see Figures 1-3).

At its simplest, the fraction of one's trips dedicated to specific activities is a classical indicator of who they may be (Lu and Pas, 1999). Persons with children spend more time at schools, while older individuals spend more time shopping. Similarly, the fraction of trips taken with modes can be indicative of sociodemographic background. In the Seattle context, those with higher incomes tend to drive more, while men tend to bike more than women (we detail more correlations in Section 5.1). Though this type of knowledge is not readily available from raw GPS data, there are now imputation algorithms that can infer multiple class trip purposes and modes with F1 scores up to 76% (Gao et al., 2024) and 92% (Merikhipour et al., 2024), respectively.

#### *4.1.1 Activity (node) features*

Let $\mathbf{x} = (x_1, \dots, x_N)$ be the frequency vector of an individual's $N$ distinct origin-destination (OD) purpose pairs, where each pair connects two activity locations (e.g., home, gym, school). The total trip count is $T = \sum_i x_i$. To quantify how evenly trips are distributed across these OD purpose pairs, we compute the Shannon entropy (Shannon, 1948):

$$H_t = -\sum_{i=1}^{N} \frac{x_i}{T} \log_2 \left(\frac{x_i}{T}\right), \tag{1}$$

which is always non-negative and upper-bounded by $\log_2 N$. High entropy indicates a diverse and balanced use of OD purpose pairs, while low entropy reflects a concentration of trips on just a few repeated purposes. While Shannon entropy measures the diversity of trip distribution, the Gini coefficient captures the degree of inequality in how frequently OD purpose pairs are used. Let the edge counts be sorted in non-decreasing order, $x_{(1)} \leq \cdots \leq x_{(N)}$, and define the cumulative trip count to rank $k$ as $C_k = \sum_{i \leq k} x_{(i)}$. Then the Gini coefficient (Dorfman, 1979) is:

$$G = 1 + \frac{1}{N} - \frac{2}{NT} \sum_{k=1}^{N} C_k. \tag{2}$$

A higher Gini indicates that a small number of purposes account for most trips, while a lower Gini suggests more equitable use across destinations. This reflects the extent to which an individual's travel is exploratory versus habitual, connecting to classical notions of travel regularity (Alessandretti et al., 2018; Kitamura and Van Der Hoorn, 1987).

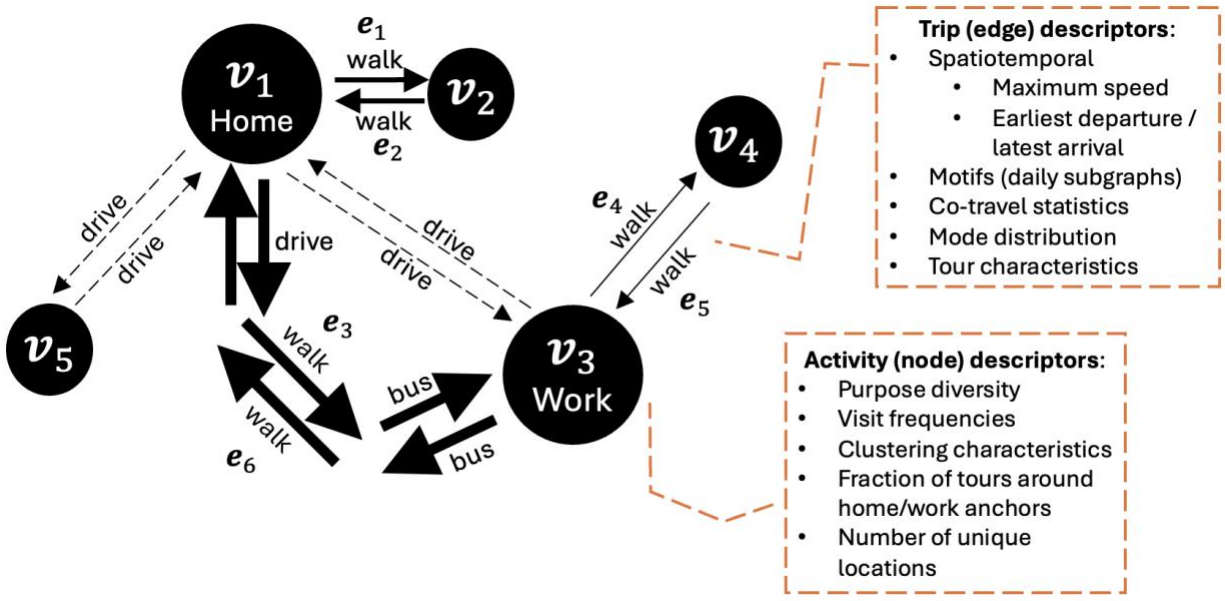


**Figure 1 Example daily mobility graph where the edges are chronologically numbered. Width of trip arrows corresponds to frequency of visits over observation period. Dashed arrows denote known trips that are not observed on this day.**

To quantify global cohesion in the mobility graph, we compute the *global clustering coefficient* $C_{\text{glob}}$ (Opsahl and Panzarasa, 2009). This metric captures the tendency for travel purposes to be mutually connected through sequences of trips, forming tightly knit triangular structures. Let $\tau_\Delta$ denote the number of closed triplets (triangles) and $\tau_\wedge$ the number of connected triplets (wedges). Then:

$$C_{\text{glob}} = \frac{3\,\tau_\Delta}{\tau_\wedge}\,. \tag{3}$$

In our context, where nodes are activity purposes and edges are trips, high clustering implies that individuals frequently travel between triplets of destinations in looping patterns (e.g., home to gym to store to home), rather than taking out-and-back trips. This metric overlaps with the concept of trip chaining, where multiple activities are linked into a single tour rather than occurring as separate out-and-back trips from a primary anchor (e.g., home or work) (Ellegard et al., 1977; Hanson, 1980). To complement the global clustering coefficient, which reflects overall network cohesion, we also compute the *mean local clustering coefficient* $\bar{c}$ *(Kaiser, 2008)*. This measure averages each node's local transitivity, placing greater emphasis on peripheral or low-degree destinations:

$$\bar{c} = \frac{1}{N}\sum_{v=1}^{N}\frac{2t_v}{k_v(k_v-1)}, \tag{4}$$

where $k_v$ is the degree of node $v \in V(G)$ (i.e., the number of other destinations directly connected to it), and $t_v$ is the number of triangles that include $v$. In mobility terms, a high $\bar{c}$ indicates that even less frequently visited destinations are embedded in tightly connected clusters. For example, this may suggest that auxiliary stops (e.g., a coffee shop, child's school) are routinely integrated into a larger, cohesive tour structure rather than occurring in isolation.

*4.1.2 Trip (edge) features*

To characterize the edge layer of each mobility graph we compute three non-overlapping subfamilies of

descriptors: spatiotemporal characteristics, daily travel motifs, and social co-travel mix. Let $n_{\text{trips}}$ and $n_{\text{tour}}$ denote, respectively, the number of observed trips and the number of closed tours (roundtrips anchored at either home or work) accumulated over the study horizon for a given individual. On the spatiotemporal side, we compute the fraction of trips taken during peak hours ($f_{\text{rush}}$), defined as 7–9 a.m. or 4–6 p.m.) and the fraction taken on weekends ($f_{\text{weekend}}$). We also extract the earliest, average, and latest departure times across all observed days. In addition, we calculate the average and maximum values for trip duration, speed, and distance over the study period.

The heterogeneity of daily mobility behavior tends to boil down to a handful of unique patterns, or "motifs" (Schneider et al., 2013). Inspired by the success attained by (Wu et al., 2019), we derive motif counts after collapsing consecutive duplicate travel purposes (e.g., *home → home → store* becomes *home → store*). The resulting sequence is parsed into one or more motifs drawn from the canonical set {single-no-return, out-and-back, chain, single-cycle, double-cycle, cycle-chain} (see Figure 2). Let $m_j$ be the count of motifs of type $j$ accumulated over all observed days and $M = \sum_j m_j$ the individual's total motif count. We retain the motif fractions $f_j = m_j/M$ together with the motif entropy $H_m = -\sum_j \frac{m_j}{M} \log_2\left(\frac{m_j}{M}\right)$ which summarizes how evenly the six patterns are observed.

Trips can also be composed of multiple modes, the use of which can correlate with income and age. Let $\mathbb{I}[\text{mm}]$ be the indicator that a given tour involves at least two distinct transport modes. The *multi-modal fraction* is then

$$f_{\text{mm}} = \frac{\mathbb{I}[\text{mm}]}{n_{\text{tour}}}. \tag{5}$$

A higher value reflects broader access to, or preference for, heterogeneous mode combinations (e.g., walk-bus-walk), which we find is positively correlated with income in the Seattle context (see Figure 5).

Finally, whether one travels with companions correlates strongly with multiple demographic labels. Thus, we tag each trip as *solo*, *with household companion(s)*, or *with non-household companion(s)*. Let $n_{\text{solo}}$, $n_{\text{hh}}$, $n_{\text{nonhh}}$ be the respective counts. We define

$$f_{\text{solo}} = \frac{n_{\text{solo}}}{n_{\text{trips}}}, \qquad f_{\text{hh}} = \frac{n_{\text{hh}}}{n_{\text{trips}}}, \qquad f_{\text{nonhh}} = \frac{n_{\text{nonhh}}}{n_{\text{trips}}}, \tag{6}$$

and composite fraction with companions $f_{\text{comp}} = 1 - f_{\text{solo}}$. These metrics embed information on household structure, caregiving roles, and wider social engagement. An illustrative computation for one travel day appears in Figure 3. Unlike purpose and mode, co-travel composition is not derived in our GPS-based pipeline used for Section 5.4. Reliable identification of whether multiple individuals are traveling together would require cross-device linkage or high-frequency synchronized sampling, which is generally infeasible under standard privacy protections and typical passive-data collection regimes. As such, co-travel features should be interpreted as survey-grade attributes that are unlikely to be directly available in most passive GPS datasets.

### 4.2 Probabilistic scoring and calibration for categorical targets

We next tackle the retention of imputation uncertainty in supervised multi-class classification. Here, we try to answer two methodological questions: (1) How can a model provide both class probabilities and a principled measure of confidence? (2) Which evaluation criteria reward not only accuracy but also well-calibrated uncertainty? We begin with definitions and then describe metrics and visual tools to understand the output of various models.

*4.2.1 Definitions*

Let $Y \in \{1, \ldots, K\}$ denote the sociodemographic label and $X$ the mobility graph descriptors, both drawn

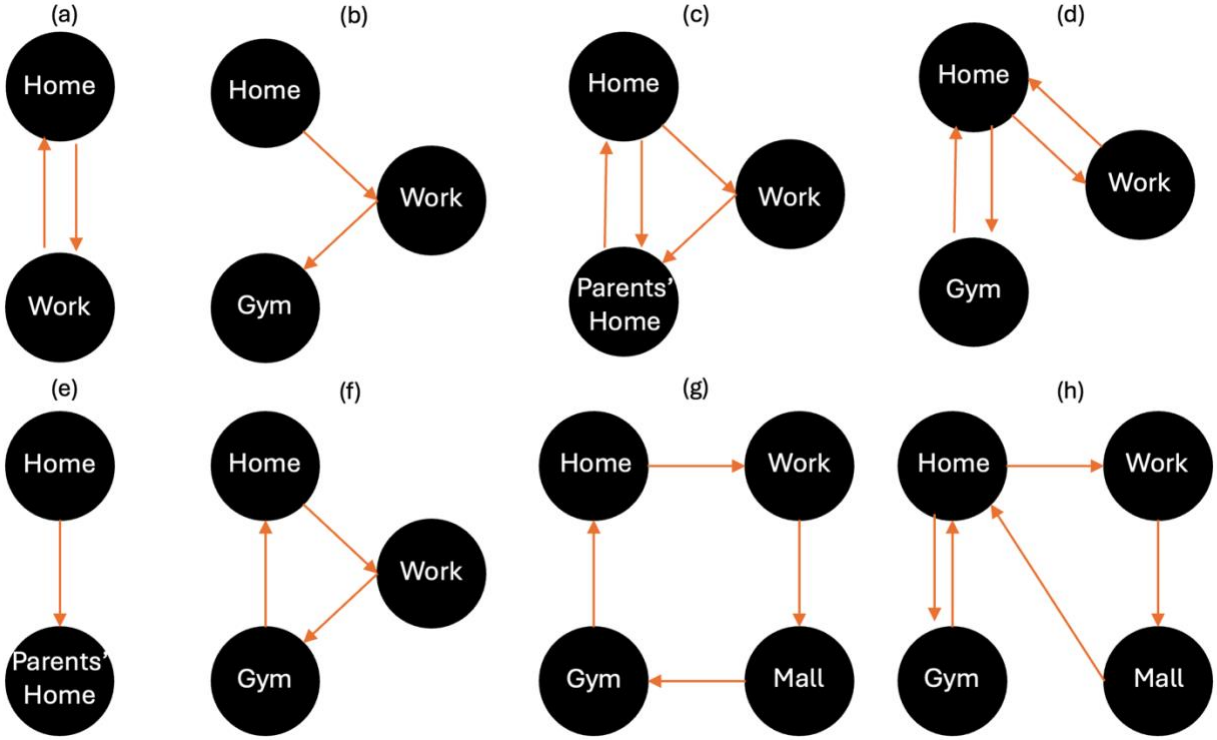


**Figure 2 Illustration of daily travel motifs (Schneider et al., 2013; Wu et al., 2019); (a) Out-and-back; (b) Chain; (c) Cycle-chain; (d, h) Double-cycle; (e) Single-no-return; (f, g) Single-Cycle**

from the ground truth joint distribution $\pi(X, Y) = \pi(Y \mid X)\pi(X)$. A model $m$ outputs $m(X) = (\hat{Y}, \hat{P})$, where $\hat{Y}$ is the predicted class and $\hat{P}$ the associated confidence (e.g., probability of correctness). We would like calibrated confidence estimates, meaning that if $m$ outputs a confidence of 0.6 for 10 predictions, roughly 6 should be correct. Formally, *perfect calibration* is defined as

$$\Pr(\hat{Y} = Y \mid \hat{P} = p) = p \quad \text{for all} \quad p \in [0,1], \tag{7}$$

where the probability is over the joint distribution. A lack of calibration can lead to biased share estimates and misleading uncertainty, whereas well-calibrated models ensure that confidence values are useful and interpretable. However, the probability above cannot be computed directly from finite samples, as $\hat{P}$ is a continuous random variable. We therefore rely on empirical approximations to estimate calibration.

*4.2.2 Metrics and reliability diagrams*

We evaluate predictions with complementary measures of separability, probability quality, and calibration. Top 1 accuracy is the fraction of test cases for which the class with highest predicted probability coincides with the true label, which is interpretable as “percentage correct”. While we acknowledge that "hard" classification via the argmax operator can be reductive in behavioral choice modeling, we justify its use here as a standard benchmark for pattern recognition, where the objective is to recover a fixed sociodemographic attribute rather than to predict a future choice.

However, because Top-1 accuracy is sensitive to class imbalance and agnostic to the quality of the full probability vector, we treat it only as a baseline. To assess separability independent of any fixed threshold, we also report the area under the ROC curve (AUROC), which measures how often the model ranks the true class above competing classes (0.5 = random, 1 = perfect). For multiclass tasks, we compute the one-against-rest macro-AUROC to ensure equal contribution from each class (Hand and Till, 2001). AUROC generalizes accuracy by integrating over all possible confidence thresholds, but it still does not assess probability calibration. A model can have high AUROC while producing poorly calibrated probabilities. Thus, we report AUROC alongside metrics that reward well-calibrated probability estimates to evaluate uncertainty quality.

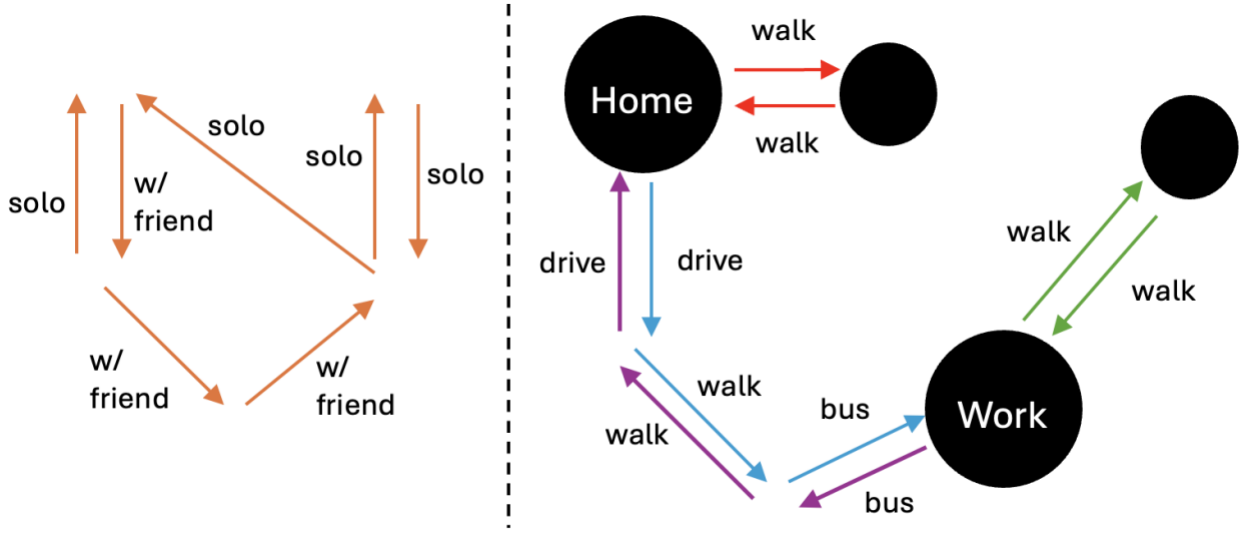


**Figure 3 Illustration of selected metrics; (left) example travel day. In this case, $n_{\text{trips}} = 7$ and $f_{\text{comp}} = \frac{3}{7}$; (right) each color denotes a tour to/from the anchors. In this case, $n_{\text{tour}} = 4$ and $f_{\text{mm}} = 2/4$.**

One such metric is the negative log-likelihood (NLL), a standard measure of a probabilistic model's quality (Hastie et al., 2001). It is also commonly called the cross-entropy loss in the context of deep learning (LeCun et al., 2015). Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ as inputs and $K$ classes, a probabilistic classifier specifies a categorical distribution $q$, where each sample has predicted class probabilities $\hat{\mathbf{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})$ requiring $\sum_k \hat{p}_{ik} = 1$. The likelihood of the observed label $y_i$ under the model for input $x_i$ is the scalar $q(y_i|\mathbf{x}_i) = \hat{p}_{i,y_i}$. Then, the NLL averages the negative log of those probabilities:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log\big(q(y_i|\mathbf{x}_i)\big) = -\frac{1}{N}\sum_{i=1}^{N} \log \hat{p}_{i,y_i}\,. \tag{8}$$

NLL is minimized when the predicted probabilities match the true conditional distribution (Gneiting and Raftery, 2007). Intuitively, NLL rewards putting high probability on the correct class and heavily penalizes overconfident mistakes.

On the other hand, the Expected Calibration Error (ECE) summarizes how well confidences match accuracies. For each sample $i$, let $\hat{y}_i = \text{argmax}_k\, \hat{p}_{ik}$ be the predicted label and the confidence of the prediction as the highest probability class $\hat{p}_i = \max_k \hat{p}_{ik}$. To derive ECE, we partition the predictions into $M$ confidence bins (each of size $1/M$). Let $B_m$ be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The accuracy and average confidence within bin $B_m$ are

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \mathbf{1}(\hat{y}_i - y_i), \tag{9}$$

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \hat{p}_i, \tag{10}$$

where $y_i$ is the actual class label for sample $i$. Given these definitions, the expected calibration error is the weighted average of the absolute accuracy-confidence gaps (Naeini et al., 2015):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \tag{11}$$

The difference between accuracy and confidence for a given bin is called the "calibration gap". Thus, a perfectly calibrated model attains $\mathrm{ECE} = 0$. In our experiments, we pre-specify $M = 15$ and use equal-width bins. One related diagnostic tool we leverage include reliability diagrams (e.g. shown in Figure 7), which are visual representations of model calibration. These diagrams plot expected sample accuracy (Eq. 9) as a function of confidence (Eq. 10). If the model is perfectly calibrated, then the diagram should plot the identity function. Points below the diagonal indicate over-confidence, while those above indicate under-confidence. Note that reliability diagrams do not display the proportion of sample in a given bin, and thus cannot be used to estimate how many samples are calibrated.

### 4.3 Multitask Learning

Multitask (MT) learning improves generalization by *inductive transfer*, the idea that learning several related tasks together leads to better performance than learning each one in isolation (Caruana, 1997). In the classical "hard parameter sharing" formulation, tasks share part of the model's parameters, which acts as a data-dependent regularizer that limits the effective hypothesis space. When tasks are related, this sharing reduces sample complexity: fewer labeled examples per task are needed to reach a given level of accuracy (Baxter, 2000).

Let $T$ tasks be indexed by $t \in \{1, \ldots, T\}$, where each task provides samples $(\mathbf{X}, \mathbf{y}_t) \sim \mathcal{D}_t$ and a task specific loss $\ell_t$. We learn a shared representation $h = g(\mathbf{X}; \boldsymbol{\theta})$ and task-specific predictors $f_t(h; \boldsymbol{\phi}_t)$ by minimizing the weighted multi-task risk

$$\min_{\theta, \{\phi_t\}} \sum_{t=1}^{T} w_t \mathbb{E}_{(\mathbf{X}, \mathbf{y}_t) \sim \mathcal{D}_t} [\ell_t(f_t(g(\mathbf{X}; \boldsymbol{\theta}); \boldsymbol{\phi}_t), \boldsymbol{y}_t)] + \Omega(\boldsymbol{\theta}, \boldsymbol{\phi}_t), \tag{14}$$

where $w_t$ are task weights, $\boldsymbol{\theta}$ are the shared parameters of the representation $h$, $\boldsymbol{\phi}_t$ are the task-specific parameters of head $f_t$, and $\Omega$ is standard parameter regularization. In practice we optimize the empirical version with mini-batches sampled from the joint dataset and backpropagate the sum of per-task losses; the shared parameters $\boldsymbol{\theta}$ receive gradients from all tasks and thereby encode common structure.

In this study, we adopt hard parameter sharing in a feed-forward network (shown in Figure 4). A shared trunk $g$ maps mobility features to a latent representation via three connected layers with rectified linear units (ReLU) and dropout. ReLU enables the network to learn complex nonlinear functions by zeroing out negative inputs, while dropout improves generalization by randomly disabling activations during training and encouraging distributed representations (LeCun et al., 2015). Four multiclass task heads $f_t$ branch from this representation to predict age, gender, income, and the number of children. This constitutes a genuine MT setup because (i) all tasks update the shared trunk during training and (ii) only a small, task-specific set of parameters is unique to each head. The training objective is the weighted sum of task losses, in which we use equal weights for each task.

The mobility descriptors given in 4.1 may capture latent routines that are jointly informative for several sociodemographic attributes. Thus, sharing this representation across the learning tasks pools statistical

evidence, which improves sample efficiency for sparsely labeled or imbalanced tasks, and regularizes confidence so that probabilities remain conservative under modest distributional shifts. Consistent with the theory above, our experiments show that the shared-trunk model often matches or exceeds single-task baselines in AUROC while delivering lower NLL and ECE, particularly when training data are limited, or the test set differs from the training set.

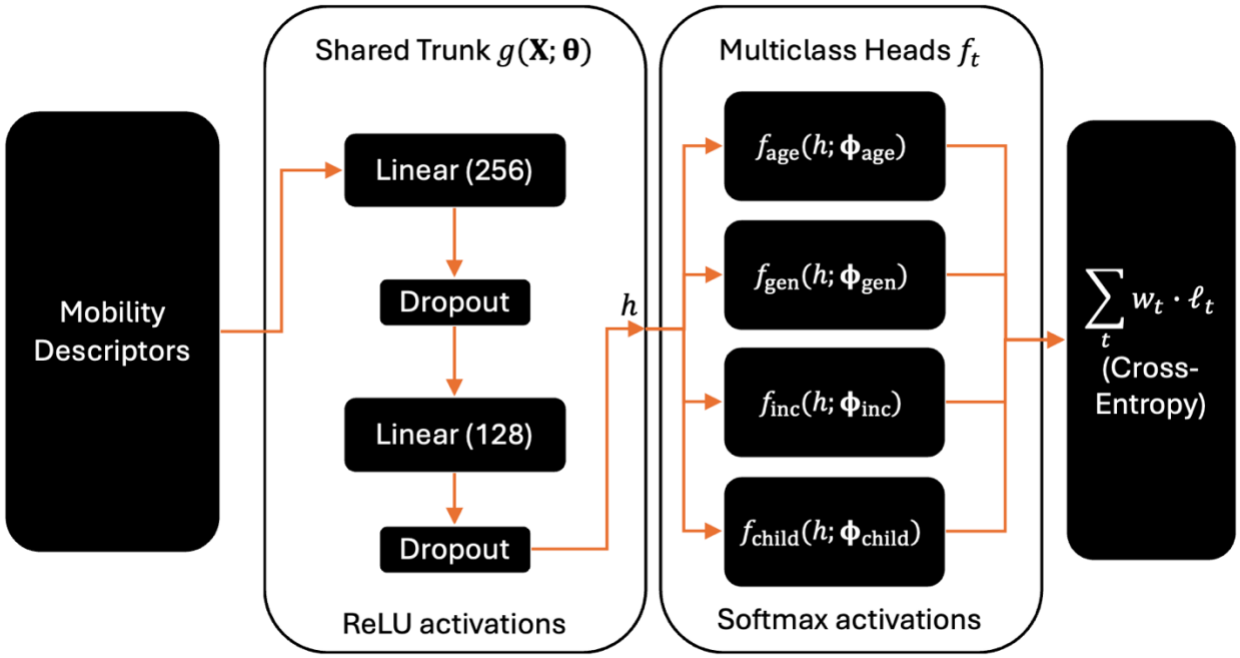


**Figure 4 Shared-trunk multitask architecture. Mobility descriptors are mapped to a shared representation $\boldsymbol{h = g(\mathbf{X};\boldsymbol{\theta})}$ by a two-layer feed-forward network with ReLU activations and dropout. Four task-specific heads $\boldsymbol{f_t}$ produce class probabilities via softmax. Training minimizes the cross-entropy loss $\boldsymbol{\sum_t w_t \cdot \ell_t}$ where we set the weights to be equal.**

## 5. EXPERIMENTS

We showcase our results. In 5.1, we analyze correlations between our descriptors and demographic targets, which fit well-known tropes in the literature. 5.2 assesses the extent to which our features improve the predictability of demographics and model calibration. In 5.3, we demonstrate the value of MTL, which improves outcomes in data scarce regimes or on test sets that differ from the training set. In 5.4, we assess how differences between variables derived from survey data and passively-inferred attributes (i.e., from GPS data) affect downstream performance, using a PSRC subsample for which both raw GPS trajectories and HTS ground-truth records are available.

### 5.1 Linkages between mobility descriptors and sociodemographics

Figure 5 shows the largest magnitude correlations between selected sociodemographics and our feature set. Age is most strongly related to the share of school-purpose travel ($\rho \approx -0.40$) and the drive-alone fraction ($\rho \approx +0.25$), with younger travelers exhibiting more bus use and group car travel. Income is negatively associated with the share of shopping trips ($\rho < 0$) but positively related to trip speed, distance, and car use, consistent with longer, faster trips among higher-income travelers. Interestingly, higher income also correlates with greater transit use, likely reflecting the prevalence of white-collar commuters in Seattle who rely on transit for downtown access. Gender effects are modest ($|\rho| \leq 0.08$) but systematic: men tend to bike more, travel solo, and have a higher fraction of their trips be work-related, while women have more complex tours, more errand/appointment trips, and more shopping trips. By contrast, the number of children shows much stronger signal (up to $|\rho| \approx 0.5$): households with more children are characterized by more carpooling (3+ in vehicle) and higher shares of school-bus, school, and escort travel, alongside lower prevalence of solo driving, transit, and walking, fewer work-anchored tours, shorter trip durations, and fewer shopping trips. These patterns reflect the child-serving, time-constrained nature of family travel.

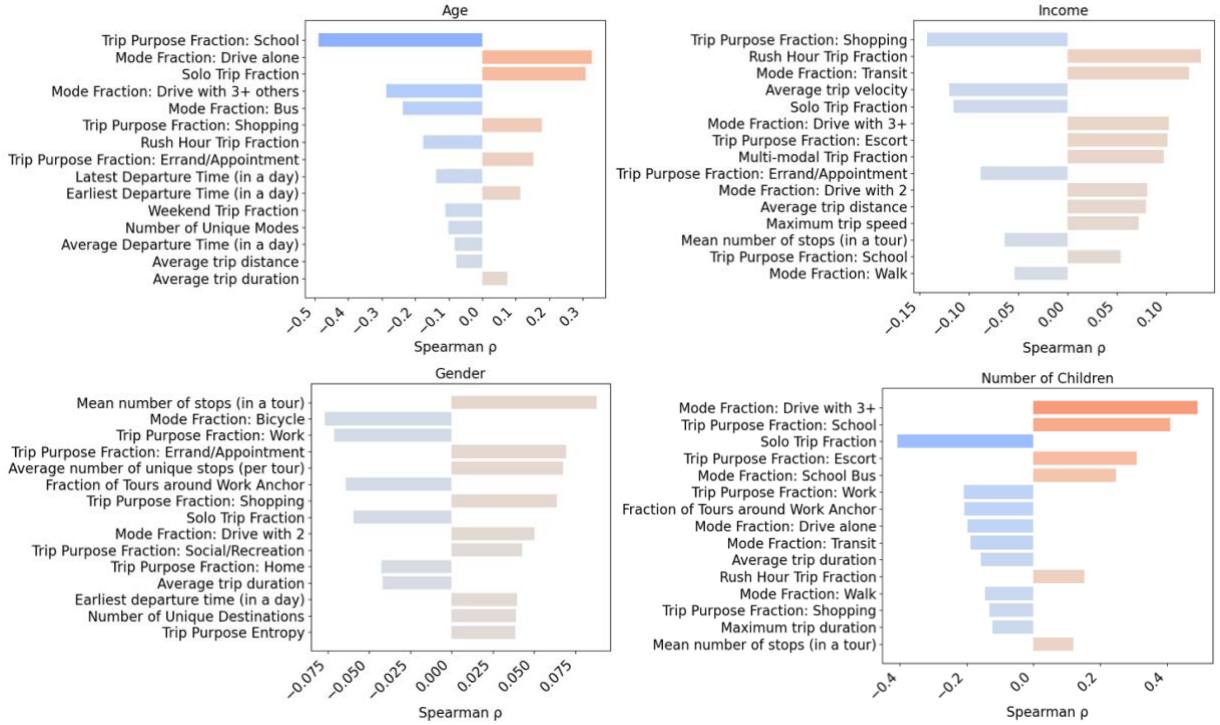


**Figure 5 Largest magnitude spearman rank correlations (all $\rho < 0.001$) between mobility descriptors and demographics. (top left): age; (top right): household income; (bottom left) gender; (bottom right) number of children. Bars to the left indicate negative associations;**

**to the right, positive.**

**Table 2 Selected linear regression models**

| Dependent Variable | Multi-modal Fraction | | Fraction with Companions | | Avg. Local Clustering Coeff. | | Out-and-back Fraction (motif) | |
|---|---|---|---|---|---|---|---|---|
| Model / Ind. Variable | M1 | M2 | M1 | M2 | M1 | M2 | M1 | M2 |
| Intercept | 0.50** | 0.53** | 0.46** | 0.61** | 0.35** | 0.35** | 0.42** | 0.41** |
| Age | 0.02** | 0.01** | -0.08** | -0.12** | -0.00 | 0.00 | -0.02** | -0.02** |
| Income | 0.03** | 0.03** | 0.01** | 0.02** | -0.00* | -0.01* | -0.00 | -0.01 |
| Gender | 0.01 | 0.01 | 0.05** | 0.06** | 0.02** | 0.02** | -0.02** | -0.02** |
| Household Size | 0.03** | -- | 0.12** | -- | -0.01 | -- | -0.01 | -- |
| $R^2$ | 0.01 | 0.01 | 0.22 | 0.16 | 0.002 | 0.001 | 0.004 | 0.003 |

**, and * signify a p-value less than 0.01, and 0.05, respectively.

Table 2 shows the results of linear regression models with the sociodemographic attributes as the independent variables and our features as the dependent variables. Due to high negative correlation between age and household size (older respondents tend to live in smaller households as children move out), we try two variants of each model. Multimodality rises with age, income, and number of children, but the explained variance is small. The fraction with companions shows the clearest sociodemographic signal: it decreases with age and increases with income, being female, and number of children. By contrast, average local clustering is only weakly related to demographics, tending to be lower at higher incomes and slightly higher among women, with small effects. Out-and-back motif fractions decline with age and among women. Overall, co-travel patterns carry the most sociodemographic information in these linear models, while clustering and motif shares add subtle signals.

The linear regression models in Table 2 also show that while sociodemographic attributes are statistically significant predictors of mobility behavior ($p < 0.01$), the explained variance ($R^2$) remains low. These low $R^2$ values reflect the inherent difficulty of sociodemographic inference: human mobility is influenced by a myriad of factors beyond demographics, leading to weak individual signals. However, the robustness of these associations (even if they explain only a small fraction of the total variance) provides the necessary foundation for the more complex, non-linear models used in the subsequent sections, which aggregate these marginal signals to achieve higher predictive accuracy.

### 5.2 Uplift of mobility graph features on predictability

We evaluated the incremental predictive value of mobility features using nested feature sets that comprise of classical variables (C), spatiotemporal attributes (ST), diversity-related metrics (D), daily motifs (M), and co-travel statistics (CT), all discussed in Section 4.1. Classical (C) covariates include purpose shares (e.g., work, shopping, leisure), mode shares (drive, transit, walk), and simple tour statistics (e.g., number of tours and the share anchored at home/work). Spatiotemporal (ST) adds departure and arrival timing (e.g., first departure, latest arrival), trip durations, shares by period (peak/off-peak, weekend), and basic speeds and distances. Diversity (D) captures how spread-out travel is across activities, modes, and origin-destination pairs (e.g., purpose entropy, the fraction of multi-modal tours, and triadic closure among destinations). Motifs (M) comprise the fractions of canonical day-level patterns (out-and-back, chains, cycles, etc.) together with a summary of their evenness. Co-travel (CT) records with whom trips are taken: the shares of solo, with-household, and with-non-household trips, and the overall accompanied fraction. In our setup, each set builds on the previous one--for example, the +ST set includes all classical variables plus spatiotemporal attributes, while the +CT set includes all preceding groups and adds co-travel statistics.

All experiments followed the same data protocol, which had two evaluation splits: on the *pooled distribution* split, we created a 70/10/20 train/validation/test partition on the combined waves of the PSRC survey, assessing the model's ability to identify stable behavioral "signatures". On the *cross-temporal* split, we trained and validated on the 2017 and 2019 waves and tested on the 2021 and 2023 waves, periods with documented shifts in mobility behavior due to the aftermath of COVID-19 (see differences in wave composition in Table 1). This allowed us to assess the model's ability to generalize to a future period of documented behavioral shift without having any prior access to that period's specific data distribution. For both splits, we used five-fold multilabel cross-validation on the pooled training and validation sets, with final performance reported on the fixed test set (either a random 20% holdout or the 2023 wave, respectively).

We evaluated performance on top-1 accuracy, area under the receiver operating curve (AUROC), negative log-likelihood (NLL), and expected calibration error (ECE), as defined previously. Figure 6 Marginal changes in top-1 accuracy (higher = better), AUROC (higher = better), NLL (lower = better), ECE (lower = better) as more features are added. (top four rows) Pooled distribution split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data presents the main results for the multi-task DNN, which is our primary model due to its compatibility with shared representation learning. To assess the robustness of the proposed feature sets across model types, we additionally evaluated three other classifiers: random forests (RF), gradient boosted machines (GBM), and support vector machines (SVM). Results for these models are reported in the Appendix (Tables A3-A10), and show that the proposed descriptors broadly improve separability and likelihood across models.

However, the most comprehensive feature sets (+M or +CT) do not always yield the best performance—approximately 28% (9 out of 32) of the top metric values (shown in red) come from simpler sets. This pattern is most pronounced under cross-temporal evaluation, where training is limited to earlier waves (2017/2019), resulting in a substantially smaller effective sample size. In this regime, richer feature sets introduce correlated covariates whose benefits are harder to estimate reliably, leading to diminishing returns and occasional variance-driven performance degradation. By contrast, under the pooled distribution split (where the training sample is larger and the distributions are more aligned), the more expressive feature sets tend to perform more consistently well. Practically, this suggests that simpler representations may be preferable in data-limited or strongly out-of-distribution settings, while richer behavioral descriptors become more advantageous as sample size increases and distributional mismatches decrease.

For the multi-task DNN shown in Figure 6, we note the following observations of interest. First, extra covariates tend to help more under the *overall* split than the *cross-temporal* split. This suggests that some of the added features may capture time-specific patterns that do not generalize well when the test data come from a different distribution. In general, while richer feature sets can increase expressiveness, they may also introduce redundancy or collinearity, amplifying variance without adding meaningful new signal—especially when model capacity is not properly regularized. This effect is not uniform across models. As shown in Tables A3-A10, gradient-boosted trees (GBMs) appear less prone to overfitting under the cross-temporal split, possibly due to their implicit regularization and ability to ignore noisy or uninformative splits.

A second trend relates to model calibration. The expected calibration error (ECE) does not consistently improve with feature richness. In two of the four tasks, the most comprehensive set (+CT) yields the best-calibrated predictions, but in the others, simpler feature sets perform better. GBMs again stand out, often producing the lowest ECE scores overall, which may reflect their ability to learn conservative margins or resist overconfidence in low-signal settings. These findings highlight that while added features can increase model expressiveness, their value depends on stability across contexts and their interaction with model calibration.

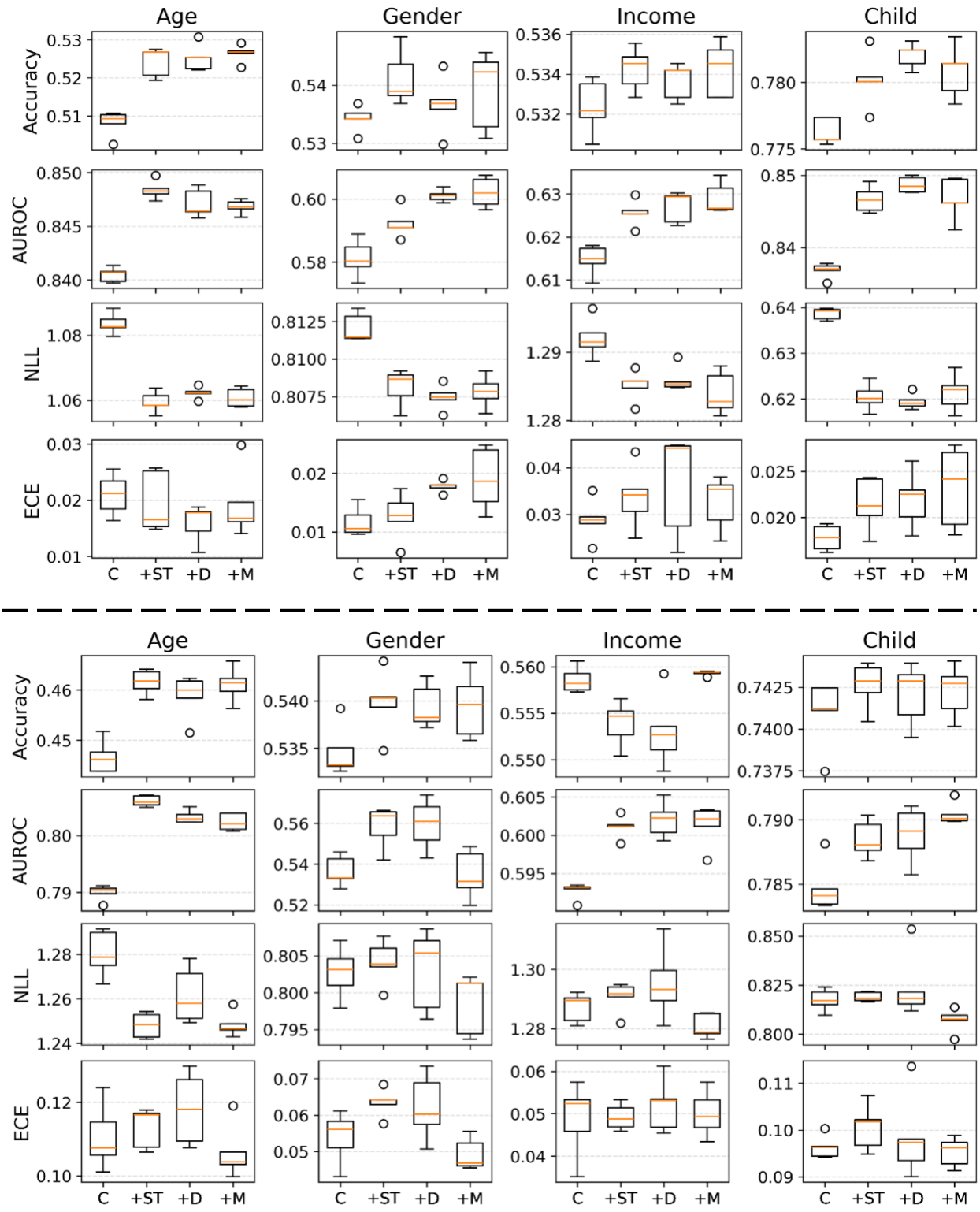


**Figure 6 Marginal changes in top-1 accuracy (higher = better), AUROC (higher = better), NLL (lower = better), ECE (lower = better) as more features are added. (top four rows) Pooled distribution split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data**

Finally, the proposed descriptors generally improve both class separability (AUROC) and model fit (NLL). Under the pooled distribution split, gains tend to increase with each feature group, with the +CT set frequently yielding the best performance and rarely underperforming simpler sets. In the cross-temporal setting, improvements are more muted and task-dependent. For Age, +CT performs best for RF and GBM and is competitive for DNNs. For Number of Children, +CT achieves the highest AUROC, while +D yields the best NLL. Results for Household Income and Gender are more variable, with no single feature set dominating across metrics or models. These patterns suggest that feature utility is context- and task-dependent, with diminishing returns or instability emerging under distribution shift.

Reliability diagrams provide hints towards clarifying the variability in ECE performance. Figure 7 highlights that, under several settings (e.g., Age in the pooled distribution split and Number of Children under the cross-temporal split), the empirical accuracy within certain confidence bins lies well above the diagonal, indicating under-confidence: the model predicts correctly more often than its stated probabilities would suggest. This conservatism leaves AUROC and accuracy unchanged or improved but increases ECE, which penalizes the magnitude of the accuracy–confidence gap irrespective of sign. Surprisingly, this seems to happen both when more features are added (as in the top row) *and* when features are removed (as in the bottom row).

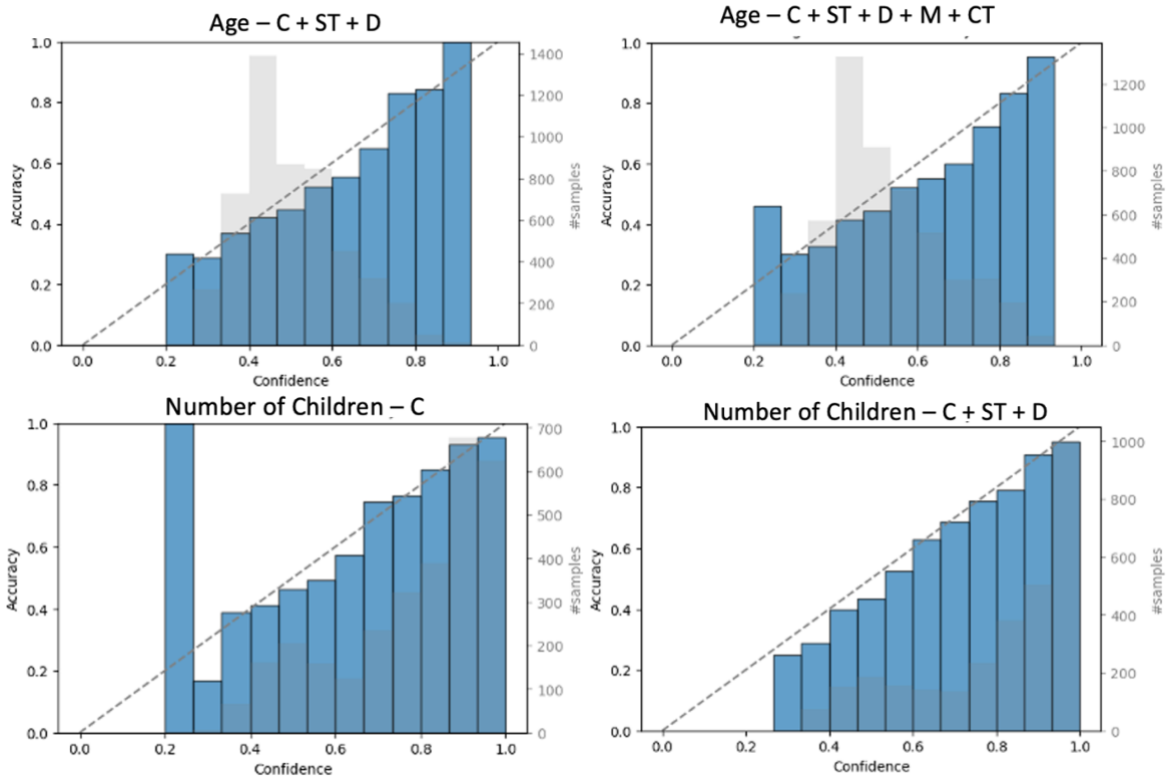


**Figure 7 Reliability diagrams for representative settings. (TOP) Pooled distribution split, Age task with C+ST+D features (left) and All features (right). (BOTTOM) Cross-temporal split, Number of Children task with C (left) and C+ST+D (right). Bars show empirical accuracy within 15 equal-width confidence bins; the dashed line is the identity (perfect calibration). Grey histograms (right axes) give the number of samples per bin. Points above (below) the diagonal indicate under- (over-) confidence.**

### 5.3 Impact of Multitask Learning

We compared a shared trunk multitask (MT) network with matched single task variants (STV) across age, gender, income, and number of children, while holding architecture, optimization, and regularization fixed (see Appendix A for full hyperparameter tuning details). To stabilize learning and reduce task interference, we optionally applied a *per-task layer normalization* module before each task head, normalizing activations separately for each prediction branch rather than sharing normalization statistics across tasks. To probe sample efficiency, we randomly subsampled the training split to fractions {1.0, 0.1 ,0.01, 0.001} and evaluated on the untouched validation and test sets. We additionally benchmarked our neural nets with a deep generative baseline (conditional variational-autoencoders, or CVAEs), providing a rigorous comparison against a paradigm that explicitly models the joint distribution of target attributes. To avoid overfitting as training data size got small, we used only the classical and spatiotemporal features.

Figure 8 displays our findings. On the pooled distribution split (top row of Figure 8), the MT network, the STVs, and the generative CVAE achieve comparable top-1 accuracies and AUROC at full data. However, as the training size fraction shrinks, the relative patterns diverge. Noticeably, both the separability and probabilistic evaluation metrics favor the MT network over both baselines as we get to the lower training data fractions. The STVs also occasionally outperform the CVAE, except for the lower data fractions in top-1 accuracy and NLL. This pattern suggests that, in our data regime, discriminative objectives are more data-efficient than the generative approach, which must learn both a latent representation and a decoder over multiple label spaces. The CVAE appears to underfit when labeled data are limited, leading to weaker class separability and higher NLL.

The advantages of multitask learning extend to settings where the test data distribution differs from that of the training data. Under the cross-temporal generalization setting (bottom row of Figure 8), the MT network outperforms the STVs and the CVAE in accuracy and AUROC across most fractions, though the advantages are uneven. The improvements in NLL are more pronounced than in the pooled distribution split, suggesting that multitask learning helps regularize against overconfidence under distribution shift. The CVAE baseline again underperforms both discriminative approaches, particularly at small and intermediate fractions, where its AUROC and NLL degrade more steeply. By pooling representational strength across targets, the discriminative MT model can better withstand changes in the marginal distribution of mobility features. This effect is especially helpful when some individual targets have relatively few informative examples in the new distribution (i.e., with rare subgroups or behaviors that shift over time).

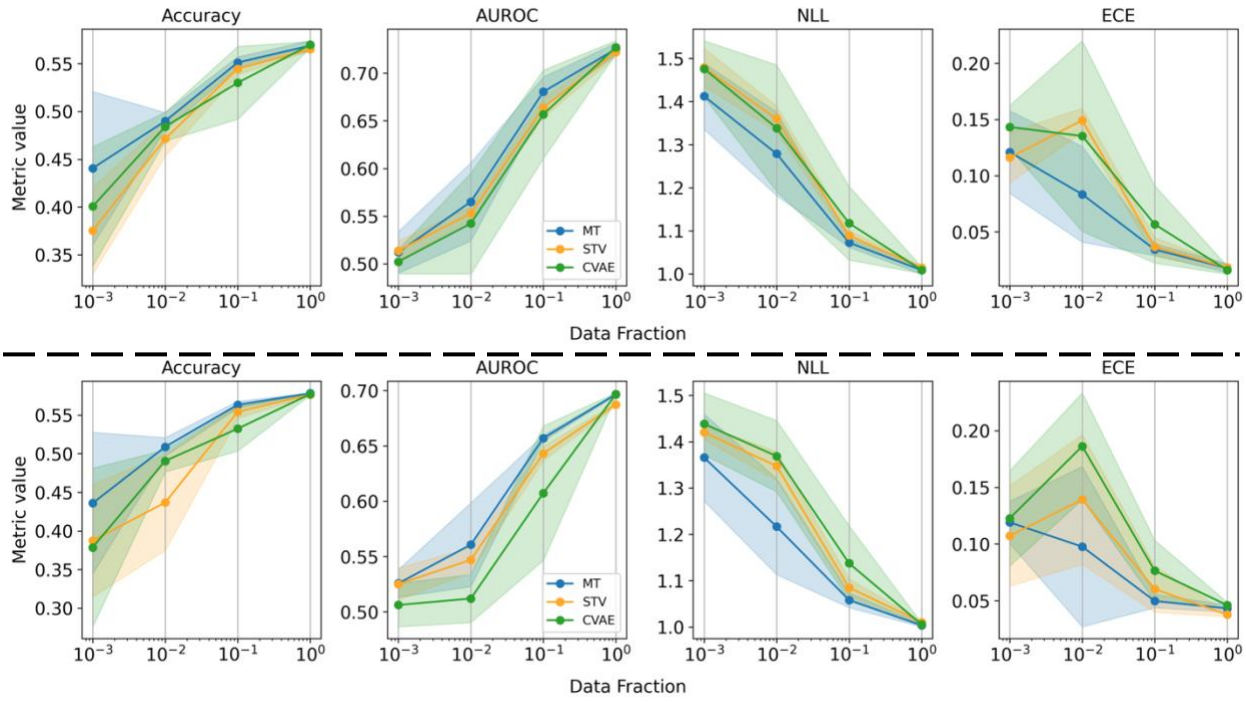


**Figure 8 Performance of the MT variant (in blue) compared to ST variants (in orange) and a CVAE (in green) at different fractions of training data. Metrics are averaged across the four tasks. (top row) Pooled distribution split; (bottom row) training with the 2017/2019 data and testing on the 2021/2023 data.**

### 5.4 Quantifying the performance gap between survey data and GPS traces

In this section, we quantify how differences between semantic labels in HTS and passively-inferred attributes from GPS data affect downstream sociodemographic prediction. Our purpose and mode inference pipeline for GPS traces relies on reverse geocoding and POI look-ups, which are inherently noisy and provide a coarser semantic resolution than the self-reported HTS trip purposes. As a result, passively inferred attributes differ from HTS labels along two key dimensions: semantic *coarseness* (fewer distinguishable categories) and systematic *bias* (misclassification due to inference errors). Appendix B details the upstream enrichment pipeline and provides empirical comparisons between behavioral descriptors derived from HTS diaries and those inferred from GPS traces, illustrating how these differences manifest in mobility patterns before they propagate into downstream demographic prediction.

To isolate and measure the impact of these two factors, we design two controlled experiments using both HTS and GPS data while holding the training set size and model architecture fixed. First, to quantify the impact of semantic scale (coarseness), we construct an HTS subset in which the original 14 trip purpose categories are merged into the same 7 broader categories used in the GPS inference pipeline. This version preserves the true HTS labels but matches the coarser semantic resolution of GPS-derived purposes, thereby isolating the effect of reduced categorical detail without introducing label noise. Second, to quantify the impact of label bias (misclassification), we construct another HTS subset in which the original purpose proportions for each individual are replaced with those estimated from their corresponding GPS traces (only done on HTS respondents who also had GPS subset available). These GPS-derived purposes are both noisy and coarser in scale. To ensure that differences are not driven solely by category aggregation, we also coarsen the original HTS labels to the same 7-category scheme. This setup allows us to isolate the effect of biased purpose inference while controlling for semantic scale.

Figure 9 summarizes the results (where bars show mean performance across tasks). Coarsening trip purposes from 14 to 7 categories leads to modest performance degradations: accuracy and AUROC decrease by approximately 0.003 and 0.011 on average, respectively. From a probabilistic perspective, coarsening increases NLL and ECE by about 0.017 and 0.0009, indicating slightly worse fit and calibration when semantic detail is reduced. Introducing biased purpose labels (via GPS-based inference) produces a similar deterioration in separability (AUROC) and likelihood (NLL), while calibration effects remain comparatively small. Together, these results suggest that both upstream semantic inference noise and reduced categorical resolution contribute to downstream performance loss, and there is a quantifiable empirical bound on expected degradation when deploying the framework on passive GPS data.

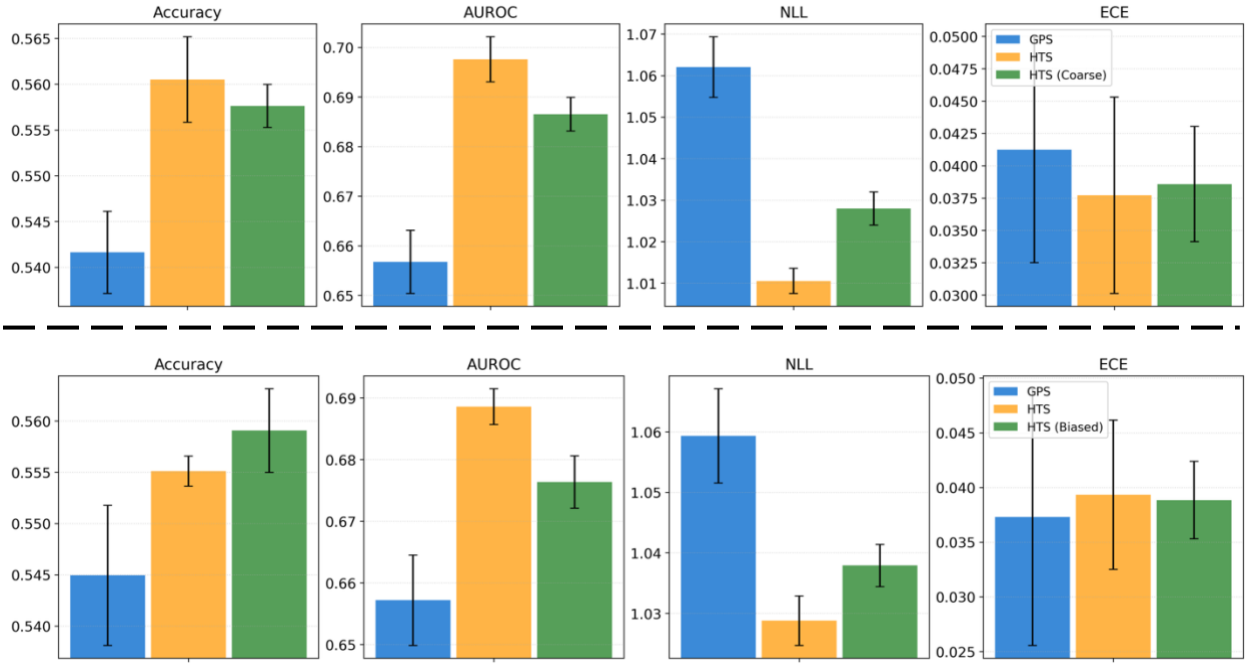


**Figure 9 Sensitivity of sociodemographic prediction to semantic coarseness and label bias in trip purposes. (top row) Effect of reducing semantic resolution from 14 HTS trip purpose categories to 7 coarser categories while preserving ground-truth labels. (bottom row) Effect of introducing biased purpose labels derived from GPS inference, with semantic scale matched to the same 7-category scheme.**

## 6. CONCLUSION

This study advances sociodemographic inference from mobility traces along three fronts: (i) it introduces a behaviorally grounded family of higher-order mobility descriptors that move beyond first-order counts to encode trip sequencing, cohesion, and social co-travel, among other characteristics; (ii) it operationalizes uncertainty-aware evaluation for multi-class prediction in this context, a dimension largely absent in prior studies, which tend to focus on point estimates without assessing the reliability of model confidence, and; (iii) it examines how a shared-trunk multitask (MT) architecture compares to matched single-task variants in data efficiency and calibration quality. Empirically, the proposed descriptors generally raise out-of-sample accuracy and lower NLL across attributes. The benefits of MT learning, however, are nuanced. While it often improves calibration and robustness under data scarcity and temporal distribution shift, these gains are not uniform across targets. In some cases—particularly for attributes with weaker behavioral overlap or higher label noise—task interference leads to mild negative transfer, causing MT to underperform specialized single-task networks. Overall, shared representations can regularize predictions

and enhance reliability, but their value depends on the degree of cross-task relatedness and the balance between shared and task-specific learning.

Importantly, the emphasis on model calibration is not only methodological but also practical. In many real-world applications, inferred sociodemographic attributes serve as inputs to downstream planning models rather than as final outputs. Well-calibrated probabilities allow analysts to propagate uncertainty from the imputation stage into subsequent analyses in a principled manner, instead of relying on single most-likely class assignments that implicitly assume perfect knowledge. For example, when enriched GPS data are used in behavioral models such as mode choice estimation, calibrated class probabilities can be incorporated probabilistically (e.g., via weighting or simulation) so that uncertainty in sociodemographic inputs is reflected in model estimation and prediction. Similarly, in accessibility and equity analyses, planners often compute metrics stratified by population groups (e.g., low-income versus high-income households). When group membership is inferred rather than observed, calibrated probabilities enable these metrics to be expressed as expected values or uncertainty intervals, providing a more transparent representation of equity outcomes.

This work has limitations that motivate future research. We do not derive features directly from raw GPS/LBS signals; instead, we rely on processed trip diaries with purpose, mode, and co-travel labels. Many descriptors could in principle be computed from passive traces (e.g., reverse-geocoded activity locations, dwell-time anchors, and mode inference from speed/acceleration and network context), but their fidelity will depend on upstream imputation methods (Gao et al., 2024; Merikhipour et al., 2024). Furthermore, our cross-temporal evaluation is confined to a single region; a natural extension is a cross-city study to assess geographic portability.

More broadly, this paper does not take a model- or architecture-centric view of predictive performance. Our emphasis is on deriving behaviorally grounded features, quantifying their contributions, and examining model uncertainty rather than pursuing architectural novelty. The only modeling variation we explore is multitask learning, chosen to test whether shared representations help under data sparsity. This is a well-established line of inquiry rather than a new algorithmic proposal. That said, there is considerable potential in moving beyond hand-crafted descriptors toward models that can automatically discover structure in large-scale mobility data. Recent advances in Transformer-based and self-supervised architectures point to promising directions, particularly for unlabeled PCM with long temporal depth, where rich behavioral regularities could be captured through pretraining (Wu et al., 2024a) and later adapted to sociodemographic inference.

## REFERENCES


Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., Baronchelli, A., 2018. Evidence for a conserved quantity in human mobility. Nat Hum Behav 2, 485–491. https://doi.org/10.1038/s41562-018-0364-x

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies, Big Data in Transportation and Traffic Engineering 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Auld, J., Mohammadian, A. (Kouros), Oliveira, M.S., Wolf, J., Bachman, W., 2015. Demographic Characterization of Anonymous Trace Travel Data. Transportation Research Record 2526, 19–28. https://doi.org/10.3141/2526-03

Baxter, J., 2000. A Model of Inductive Bias Learning. jair 12, 149–198. https://doi.org/10.1613/jair.731

Bhat, C.R., Misra, R., 1999. Discretionary activity time allocation of individuals between in-home and out-of-home and between weekdays and weekends. Transportation 26, 193–229. https://doi.org/10.1023/A:1005192230485

Bian, R., Tolford, T., Liu, S., Gangireddy, S., 2023. Lessons learned from evaluating complete streets project outcomes with emerging data sources. Transportation Planning and Technology 46, 754–772. https://doi.org/10.1080/03081060.2023.2214136

Caruana, R., 1997. Multitask Learning. Machine Learning 41–75. https://doi.org/10.1023/A:1007379606734

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Ding, S., Huang, H., Zhao, T., Fu, X., 2019. Estimating Socioeconomic Status via Temporal-Spatial Mobility Analysis - A Case Study of Smart Card Data, in: 2019 28th International Conference on Computer Communication and Networks (ICCCN). Presented at the 2019 28th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9. https://doi.org/10.1109/ICCCN.2019.8847051

Doi, S., Mizuno, T., Fujiwara, N., 2021. Estimation of socioeconomic attributes from location information. J Comput Soc Sci 4, 187–205. https://doi.org/10.1007/s42001-020-00073-w

Dorfman, R., 1979. A Formula for the Gini Coefficient. The Review of Economics and Statistics 61, 146–149. https://doi.org/10.2307/1924845

Ellegard, K., Hagerstrand, T., Lenntorp, B., 1977. Activity Organization and the Generation of Daily Travel: Two Future Alternatives. Economic Geography 53, 126. https://doi.org/10.2307/142721

Gao, L., Huang, H., Ye, J., Wang, D., 2024. Activity type detection of mobile phone data based on self-training: Application of the teacher–student cycling model. Transportation Research Part C: Emerging Technologies 161, 104550. https://doi.org/10.1016/j.trc.2024.104550

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359–378. https://doi.org/10.1198/016214506000001437

Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171–186. https://doi.org/10.1023/A:1010920819831

Hanson, S., 1980. The importance of the multi-purpose journey to work in urban travel behavior. Transportation 9, 229–248. https://doi.org/10.1007/BF00153866

Hastie, T., Friedman, J., Tibshirani, R., 2001. The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-21606-5

Iqbal, Md.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A. 'Sandy', De Montjoye, Y.-A., 2017. Improving official statistics in emerging markets using machine learning and mobile phone data. EPJ Data Sci. 6, 3. https://doi.org/10.1140/epjds/s13688-017-0099-3

Kaiser, M., 2008. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. New J. Phys. 10, 083042. https://doi.org/10.1088/1367-2630/10/8/083042

Kitamura, R., Van Der Hoorn, T., 1987. Regularity and irreversibility of weekly travel behavior. Transportation 14, 227–251. https://doi.org/10.1007/BF00837531

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lee, Y., Hickman, M., Washington, S., 2007. Household type and structure, time-use pattern, and trip-chaining behavior. Transportation Research Part A: Policy and Practice 41, 1004–1020. https://doi.org/10.1016/j.tra.2007.06.007

Li, Z., Ning, H., Jing, F., Lessani, M.N., 2024. Understanding the bias of mobile location data across spatial scales and over time: A comprehensive analysis of SafeGraph data in the United States. PLOS ONE 19, e0294430. https://doi.org/10.1371/journal.pone.0294430

Lu, X., Pas, E.I., 1999. Socio-demographics, activity participation and travel behavior. Transportation Research Part A: Policy and Practice 33, 1–18. https://doi.org/10.1016/S0965-8564(98)00020-2

McGuckin, N., Murakami, E., 1999. Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women. Transportation Research Record: Journal of the Transportation Research Board 1693, 79–85. https://doi.org/10.3141/1693-12

Merikhipour, M., Khanmohammadidoustani, S., Abbasi, M., 2024. Transportation mode detection through spatial attention-based transductive long short-term memory and off-policy feature selection. Expert Syst. Appl. 267. https://doi.org/10.1016/j.eswa.2024.126196

Mokhtarian, P.L., Chen, C., 2004. TTB or not TTB, that is the question: a review and analysis of the empirical literature on travel time (and money) budgets. Transportation Research Part A: Policy and Practice 38, 643–675. https://doi.org/10.1016/j.tra.2003.12.004

Naeini, M.P., Cooper, G., Hauskrecht, M., 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence 29. https://doi.org/10.1609/aaai.v29i1.9602

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Social Networks 31, 155–163. https://doi.org/10.1016/j.socnet.2009.02.002

Razavi, R., Xue, G., Akpan, I.J., 2024. Predicting Sociodemographic Attributes from Mobile Usage Patterns: Applications and Privacy Implications. Big Data 12, 213–228. https://doi.org/10.1089/big.2022.0182

Ryan, J., Maoh, H., Kanaroglou, P., 2009. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. Geographical Analysis 41, 181–203. https://doi.org/10.1111/j.1538-4632.2009.00750.x

Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unravelling daily human mobility motifs. Journal of The Royal Society Interface 10, 20130246.

https://doi.org/10.1098/rsif.2013.0246
Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Sheller, M., Urry, J., 2006. The New Mobilities Paradigm. Environ Plan A 38, 207–226. https://doi.org/10.1068/a37268
Solomon, A., Bar, A., Yanai, C., Shapira, B., Rokach, L., 2018. Predict Demographic Information Using Word2vec on Spatial Trajectories, in: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP '18. Association for Computing Machinery, New York, NY, USA, pp. 331–339. https://doi.org/10.1145/3209219.3209224
Ugurel, E., Wu, X., Wang, R., Lee, B.H.Y., Chen, C., 2024. Metropolitan Planning Organizations' Uses of and Needs for Big Data. Findings. https://doi.org/10.32866/001c.127143
Vo, K., 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. https://doi.org/https://doi.org/10.1016/j.trb.2024.103128
Wang, Y., Guan, X., Ugurel, E., Chen, C., Huang, S., Wang, Q.R., 2025. Exploring biases in travel behavior patterns in big passively generated mobile data from 11 U.S. cities. Journal of Transport Geography 123, 104108. https://doi.org/10.1016/j.jtrangeo.2024.104108
Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. J. R. Soc. Interface. 10, 20120986. https://doi.org/10.1098/rsif.2012.0986
Williamson, P., Birkin, M., Rees, P.H., 1998. The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records. Environ Plan A 30, 785–816. https://doi.org/10.1068/a300785
Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., Liu, Y., 2019. Inferring demographics from human trajectories and geographical context. Computers, Environment and Urban Systems 77, 101368. https://doi.org/10.1016/j.compenvurbsys.2019.101368
Wu, X., He, H., Wang, Y., Wang, Q., 2024a. Pretrained Mobility Transformer: A Foundation Model for Human Mobility. https://doi.org/10.48550/arXiv.2406.02578
Wu, X., Wang, Y., Ugurel, E., Chen, C., Huang, S., Wang, Q.R., 2024b. Location-Based Service (LBS) Data Quality Metrics and Effects on Mobility Inference. https://doi.org/10.48550/arXiv.2411.16595
Xu, F., Lin, Z., Xia, T., Guo, D., Li, Y., 2020. SUME: Semantic-enhanced Urban Mobility Network Embedding for User Demographic Inference. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4, 98:1-98:25. https://doi.org/10.1145/3411807
Yang, T., Xu, X., Guo, Q., Zhang, L., Sun, H., 2017. EV charging behaviour analysis and modelling based on mobile crowdsensing data. IET Generation, Transmission & Distribution 11, 1683–1691. https://doi.org/10.1049/iet-gtd.2016.1200
Zhang, B., Rasouli, S., Feng, T., 2024. Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach. Travel Behaviour and Society 37, 100843. https://doi.org/10.1016/j.tbs.2024.100843
Zhao, Y., Pawlak, J., Sivakumar, A., 2022. Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach. Transportation Research Part B: Methodological 155, 101–134. https://doi.org/10.1016/j.trb.2021.11.004
Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X., 2015. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15. Association for

Computing Machinery, New York, NY, USA, pp. 295–304. https://doi.org/10.1145/2684822.2685287

## APPENDIX A. EXPERIMENTAL DETAILS AND FULL MODEL RESULTS

We detail the experimental setup for the rest of the classifiers used in the two experiments discussed in the main text. Both the multi-task DNNs and their single-task counterparts were built with two hidden layers, each separated by ReLU activations, and a uniform dropout rate of \$0.3\$ between layers. For optimization, we performed a grid search over learning rates {1 x $10^{-3}$, 1 x $10^{-4}$, 5 x $10^{-5}$, 1 x $10^{-5}$}, batch sizes {16, 32, 64, 128}, and weight decay values {$10^{-3}$, $10^{-4}$, $10^{-5}$}. The best configuration, determined via validation loss, used a learning rate of 5 x $10^{-5}$, batch size of 64, and weight decay of $10^{-4}$. Models were trained for up to 200 epochs with early stopping if the validation loss did not improve for 20 consecutive epochs.

To isolate the effect of parameter sharing (MT) from raw capacity, we matched aggregate hidden width across conditions. The unified MT model used a shared trunk of 256 → 128 units, while each single-task variant (STV) used 64 → 32 units. Since there are four STVs, their combined width (4 x 64, 4 x 32) is comparable to the MT trunk (256, 128), making representational capacity roughly parity-matched while differing in whether features are learned jointly or separately. Despite the smaller per-model widths, STVs require training four independent networks (four forward/backward passes, four optimizers, four early-stopping loops) and thus incur higher wall-clock time than the single unified MT model, which amortizes the trunk compute across tasks. Empirically, STVs were consistently slower than MT across data fractions (see Figure 10).

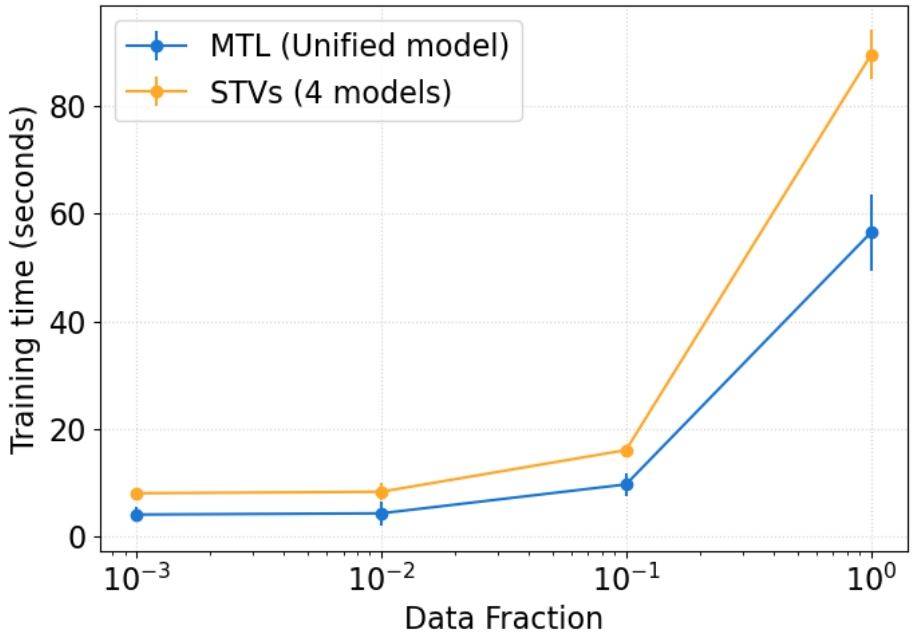


**Figure 10 Comparison of training times between the unified multi-task learning (MT) model and four separate single-task variants (STVs) across varying data fractions (log-scaled on x-axis). Despite its larger architecture, the MT model trains faster overall because its shared trunk amortizes computation across tasks, whereas STVs require four independent forward–backward passes. Error bars show the standard deviation across cross-validation folds.**

## APPENDIX B. GPS DATA ENRICHMENT AND BIAS COMPARED TO HTS

In this section, we describe the upstream processing pipeline used to enrich the subset of PSRC's HTS respondents for whom GPS trajectories are also available. Specifically, this pipeline is used to infer trip purposes and travel modes. We then quantify how the resulting behavioral descriptors differ from those derived from ground-truth HTS diaries.

Trip purposes in GPS data are inferred through a multi-step enrichment pipeline that operates on raw trajectory points. First, GPS pings are segmented into trips and stay locations using distance, speed, and temporal thresholds. Stay points are then spatially clustered using a DBSCAN algorithm with a fixed radius (200 m), which defines candidate activity locations. From these clustered stops, home and work locations are inferred using longitudinal heuristics: home is identified as the most frequent origin of first trips of the day, while work is inferred from recurrent weekday morning destinations that are distinct from home. These anchor locations are later used both as direct purpose labels and as fallbacks when POI-based labeling fails.

For non-home/work activities, trip purposes are assigned via reverse geocoding and point-of-interest (POI) matching. Each trip origin and destination is matched to nearby POIs within a spatial buffer, and POI categories are mapped to a reduced set of activity purposes. When no confident POI match is found, purposes are imputed using proximity to inferred home or work, or labeled as “other.” This process introduces multiple sources of bias: POI databases are incomplete and uneven in coverage, mixed-use areas can lead to ambiguous matches, and the fixed spatial radius may associate a stop with an incorrect venue in dense urban settings. Furthermore, the mapping from detailed POI taxonomies to a smaller purpose vocabulary reduces semantic resolution relative to the 14-category HTS scheme.

Travel mode is inferred using trajectory-derived kinematic features. Trip-level average and maximum speeds are computed from successive GPS points, and thresholds are applied to distinguish motorized from non-motorized travel. These rules are sensitive to GPS sampling frequency, signal noise, and traffic conditions; for example, congestion can make motorized trips appear slow, while sparse sampling can inflate apparent speeds. Because no direct sensing of vehicle type or transit usage is available, this step can misclassify short car trips as walking or fail to distinguish among motorized modes.

Figure 11 compares key behavioral descriptors derived from HTS and GPS-inferred data at the individual level. Daily travel motifs (top panel) show broadly similar distributions across inferred and true data, indicating that coarse sequence structures (e.g., out-and-back tours or chained trips) are reasonably preserved. Trip purpose fractions (middle panel) reveal clearer systematic differences. Home and work shares are relatively underestimated in GPS traces, while purposes like meals out, shopping, and errand-running are overestimated. Diversity and graph-based metrics (bottom panel), including entropy, Gini coefficient, number of edges, and number of unique purposes, show that inferred traces tend to exhibit slightly lower diversity and fewer distinct activity types. This is consistent with the reduced purpose vocabulary and occasional misclassification collapsing multiple activities into broader categories.

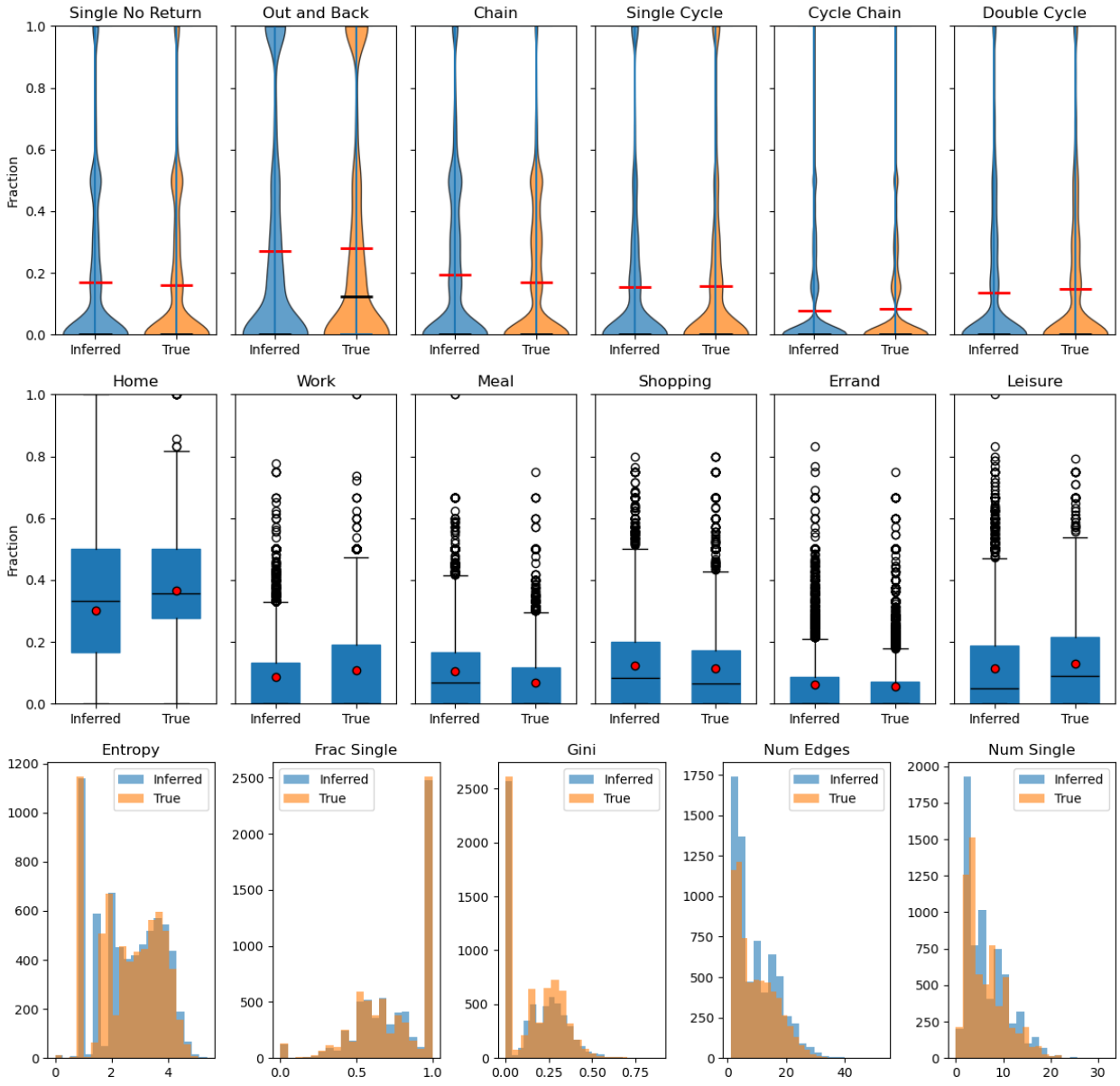


**Figure 11 Sensitivity Analysis: Comparison of behavioral descriptors derived from ground-truth HTS diaries versus noisy GPS-inferred traces. (top row) Daily motifs. (middle row) Trip purposes. (bottom row) Diversity measures**

## APPENDIX C. FULL MODEL RESULTS

For the benchmark models leveraged in the uplift experiments, we used sklearn's built-in random forest (RF) classifier, gradient boosting (GB) classifier, and C-Support Vector (SVC) classifier. For RFs and SVCs, we performed a small grid search over the number of estimators {100, 300, 500} and regularization parameter C $\in$ {0.1, 1, 10}, respectively. We selected 500 estimators for RFs and C = 1 for SVCs based on validation performance. Due to computational load, we kept the number of GB estimators at 100. We also enabled probability estimates for the SVC classifier, which slowed down convergence. The rest of this section contains the full results tables in A3 until A10.

### Pooled Distribution Split

**Table 3 Performance across feature sets and models for Age (Pooled Distribution split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.508 ± 0.003 | 0.483 ± 0.003 | 0.503 ± 0.000 | 0.504 ± 0.000 |
| | +ST | 0.524 ± 0.004 | 0.516 ± 0.004 | 0.513 ± 0.000 | 0.525 ± 0.001 |
| | +D | 0.525 ± 0.004 | 0.517 ± 0.003 | 0.515 ± 0.000 | 0.523 ± 0.000 |
| | +M | **0.526 ± 0.002** | 0.518 ± 0.003 | 0.514 ± 0.000 | 0.520 ± 0.000 |
| | +CT | 0.525 ± 0.003 | **0.520 ± 0.003** | **0.518 ± 0.000** | **0.527 ± 0.001** |
| AUROC | C | 0.840 ± 0.001 | 0.815 ± 0.000 | 0.835 ± 0.000 | 0.824 ± 0.000 |
| | +ST | 0.848 ± 0.001 | 0.834 ± 0.001 | 0.842 ± 0.000 | 0.838 ± 0.000 |
| | +D | 0.847 ± 0.001 | 0.834 ± 0.001 | 0.842 ± 0.000 | 0.838 ± 0.000 |
| | +M | 0.847 ± 0.001 | 0.834 ± 0.001 | 0.842 ± 0.000 | 0.836 ± 0.000 |
| | +CT | **0.850 ± 0.001** | **0.837 ± 0.001** | **0.843 ± 0.000** | **0.838 ± 0.000** |
| NLL | C | 1.084 ± 0.003 | 1.277 ± 0.007 | 1.099 ± 0.000 | 1.134 ± 0.000 |
| | +ST | 1.059 ± 0.003 | 1.117 ± 0.006 | 1.080 ± 0.000 | 1.096 ± 0.000 |
| | +D | 1.062 ± 0.002 | 1.118 ± 0.003 | 1.081 ± 0.000 | 1.096 ± 0.000 |
| | +M | 1.061 ± 0.003 | 1.120 ± 0.006 | 1.081 ± 0.000 | 1.099 ± 0.000 |
| | +CT | **1.048 ± 0.002** | **1.098 ± 0.003** | **1.071 ± 0.000** | **1.086 ± 0.000** |
| ECE | C | 0.021 ± 0.004 | 0.060 ± 0.004 | 0.028 ± 0.001 | 0.034 ± 0.001 |
| | +ST | 0.020 ± 0.005 | **0.029 ± 0.003** | 0.016 ± 0.001 | 0.027 ± 0.002 |
| | +D | **0.016 ± 0.003** | 0.032 ± 0.006 | 0.017 ± 0.000 | 0.022 ± 0.003 |
| | +M | 0.019 ± 0.006 | 0.034 ± 0.008 | 0.017 ± 0.000 | **0.021 ± 0.001** |
| | +CT | 0.018 ± 0.006 | 0.034 ± 0.003 | **0.016 ± 0.000** | 0.023 ± 0.002 |

**Table 4 Performance across feature sets and models for Gender (Pooled Distribution split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| **Accuracy** | C | 0.534 ± 0.002 | 0.519 ± 0.002 | 0.528 ± 0.000 | 0.530 ± 0.001 |
| | +ST | 0.541 ± 0.005 | 0.541 ± 0.002 | 0.547 ± 0.000 | 0.538 ± 0.000 |
| | +D | 0.537 ± 0.005 | 0.541 ± 0.006 | 0.543 ± 0.000 | 0.541 ± 0.000 |
| | +M | 0.539 ± 0.007 | **0.541 ± 0.004** | 0.545 ± 0.000 | 0.537 ± 0.000 |
| | +CT | **0.541 ± 0.007** | 0.540 ± 0.006 | **0.549 ± 0.000** | **0.546 ± 0.001** |
| **AUROC** | C | 0.581 ± 0.006 | 0.558 ± 0.002 | 0.588 ± 0.000 | 0.588 ± 0.001 |
| | +ST | 0.592 ± 0.005 | 0.587 ± 0.004 | 0.606 ± 0.000 | 0.588 ± 0.000 |
| | +D | 0.601 ± 0.002 | 0.591 ± 0.006 | 0.628 ± 0.000 | 0.596 ± 0.000 |
| | +M | 0.602 ± 0.005 | **0.594 ± 0.003** | **0.631 ± 0.000** | 0.608 ± 0.000 |
| | +CT | **0.611 ± 0.002** | 0.593 ± 0.007 | 0.627 ± 0.000 | **0.612 ± 0.000** |
| **NLL** | C | 0.812 ± 0.001 | 0.929 ± 0.002 | 0.813 ± 0.000 | 0.814 ± 0.000 |
| | +ST | 0.808 ± 0.001 | 0.832 ± 0.002 | 0.806 ± 0.000 | 0.810 ± 0.000 |
| | +D | 0.807 ± 0.001 | 0.831 ± 0.003 | 0.802 ± 0.000 | 0.809 ± 0.000 |
| | +M | 0.808 ± 0.001 | **0.829 ± 0.000** | 0.802 ± 0.000 | 0.808 ± 0.000 |
| | +CT | **0.805 ± 0.001** | 0.829 ± 0.002 | **0.801 ± 0.000** | **0.806 ± 0.000** |
| **ECE** | C | **0.012 ± 0.002** | 0.077 ± 0.002 | 0.019 ± 0.000 | 0.008 ± 0.001 |
| | +ST | 0.013 ± 0.004 | 0.040 ± 0.006 | 0.013 ± 0.000 | **0.005 ± 0.001** |
| | +D | 0.018 ± 0.001 | 0.041 ± 0.007 | 0.015 ± 0.000 | 0.010 ± 0.001 |
| | +M | 0.019 ± 0.005 | **0.035 ± 0.003** | 0.014 ± 0.001 | 0.006 ± 0.001 |
| | +CT | 0.021 ± 0.006 | 0.039 ± 0.004 | **0.013 ± 0.000** | 0.013 ± 0.002 |

**Table 5 Performance across feature sets and models for HH Income (Pooled Distribution split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.532 ± 0.001 | 0.528 ± 0.002 | 0.535 ± 0.000 | 0.532 ± 0.000 |
| | +ST | 0.534 ± 0.001 | 0.561 ± 0.001 | 0.531 ± 0.000 | 0.535 ± 0.001 |
| | +D | 0.534 ± 0.001 | **0.563 ± 0.001** | 0.534 ± 0.000 | 0.536 ± 0.001 |
| | +M | 0.534 ± 0.001 | 0.561 ± 0.001 | 0.532 ± 0.000 | 0.536 ± 0.001 |
| | +CT | **0.538 ± 0.000** | 0.561 ± 0.001 | **0.536 ± 0.000** | **0.540 ± 0.001** |
| AUROC | C | 0.615 ± 0.003 | 0.617 ± 0.001 | 0.612 ± 0.000 | 0.599 ± 0.001 |
| | +ST | 0.626 ± 0.003 | 0.677 ± 0.001 | 0.638 ± 0.000 | 0.626 ± 0.000 |
| | +D | 0.627 ± 0.004 | 0.678 ± 0.002 | 0.639 ± 0.000 | 0.630 ± 0.000 |
| | +M | 0.629 ± 0.004 | 0.678 ± 0.002 | 0.640 ± 0.000 | 0.631 ± 0.000 |
| | +CT | **0.639 ± 0.003** | **0.691 ± 0.001** | **0.650 ± 0.000** | **0.644 ± 0.000** |
| NLL | C | 1.292 ± 0.003 | 1.556 ± 0.013 | 1.289 ± 0.000 | 1.305 ± 0.000 |
| | +ST | 1.285 ± 0.002 | 1.216 ± 0.001 | 1.275 ± 0.000 | 1.291 ± 0.000 |
| | +D | 1.286 ± 0.002 | 1.216 ± 0.002 | 1.273 ± 0.000 | 1.288 ± 0.000 |
| | +M | 1.284 ± 0.003 | 1.217 ± 0.003 | 1.272 ± 0.000 | 1.287 ± 0.000 |
| | +CT | **1.275 ± 0.002** | **1.202 ± 0.002** | **1.263 ± 0.000** | **1.275 ± 0.000** |
| ECE | C | **0.029 ± 0.004** | 0.048 ± 0.002 | 0.029 ± 0.000 | **0.025 ± 0.001** |
| | +ST | 0.034 ± 0.007 | **0.020 ± 0.003** | **0.022 ± 0.000** | 0.041 ± 0.004 |
| | +D | 0.037 ± 0.011 | 0.022 ± 0.003 | 0.031 ± 0.001 | 0.047 ± 0.001 |
| | +M | 0.033 ± 0.006 | 0.020 ± 0.002 | 0.031 ± 0.000 | 0.047 ± 0.001 |
| | +CT | 0.035 ± 0.007 | 0.027 ± 0.004 | 0.024 ± 0.000 | 0.064 ± 0.001 |

**Table 6 Performance across feature sets and models for Number of Children (Pooled Distribution split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| **Accuracy** | C | 0.776 ± 0.001 | 0.783 ± 0.001 | 0.779 ± 0.000 | 0.774 ± 0.000 |
| | +ST | 0.780 ± 0.002 | 0.807 ± 0.001 | 0.784 ± 0.000 | 0.782 ± 0.000 |
| | +D | 0.782 ± 0.001 | 0.805 ± 0.000 | 0.783 ± 0.000 | 0.786 ± 0.000 |
| | +M | 0.781 ± 0.002 | 0.804 ± 0.001 | 0.781 ± 0.000 | 0.786 ± 0.001 |
| | +CT | **0.785 ± 0.002** | **0.808 ± 0.001** | **0.787 ± 0.000** | **0.791 ± 0.000** |
| **AUROC** | C | 0.837 ± 0.001 | 0.823 ± 0.001 | 0.842 ± 0.000 | 0.831 ± 0.000 |
| | +ST | 0.847 ± 0.002 | 0.861 ± 0.001 | 0.852 ± 0.000 | 0.844 ± 0.000 |
| | +D | 0.849 ± 0.001 | 0.862 ± 0.002 | 0.853 ± 0.000 | 0.845 ± 0.000 |
| | +M | 0.847 ± 0.003 | 0.861 ± 0.001 | 0.854 ± 0.000 | 0.844 ± 0.000 |
| | +CT | **0.854 ± 0.002** | **0.862 ± 0.001** | **0.859 ± 0.000** | **0.851 ± 0.000** |
| **NLL** | C | 0.639 ± 0.001 | 0.801 ± 0.008 | 0.634 ± 0.000 | 0.650 ± 0.000 |
| | +ST | 0.620 ± 0.003 | 0.595 ± 0.003 | 0.615 ± 0.000 | 0.625 ± 0.000 |
| | +D | 0.619 ± 0.002 | 0.597 ± 0.008 | 0.613 ± 0.000 | 0.623 ± 0.000 |
| | +M | 0.621 ± 0.004 | 0.596 ± 0.005 | 0.613 ± 0.000 | 0.623 ± 0.000 |
| | +CT | **0.610 ± 0.003** | **0.595 ± 0.010** | **0.604 ± 0.000** | **0.613 ± 0.000** |
| **ECE** | C | 0.018 ± 0.001 | **0.030 ± 0.003** | 0.018 ± 0.000 | 0.059 ± 0.001 |
| | +ST | 0.021 ± 0.003 | 0.055 ± 0.001 | 0.019 ± 0.000 | 0.045 ± 0.002 |
| | +D | 0.022 ± 0.003 | 0.057 ± 0.002 | 0.023 ± 0.000 | 0.035 ± 0.001 |
| | +M | 0.023 ± 0.004 | 0.058 ± 0.001 | 0.020 ± 0.000 | 0.038 ± 0.001 |
| | +CT | **0.016 ± 0.006** | 0.048 ± 0.002 | **0.017 ± 0.000** | **0.030 ± 0.001** |

**Cross-temporal Split**

**Table 7 Performance across feature sets and models for Age (2017/2019 train, 2021/2023 test split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.474 ± 0.002 | 0.443 ± 0.001 | 0.471 ± 0.000 | 0.469 ± 0.000 |
| | +ST | **0.488 ± 0.004** | 0.480 ± 0.002 | 0.489 ± 0.000 | 0.484 ± 0.001 |
| | +D | 0.480 ± 0.003 | 0.479 ± 0.004 | 0.485 ± 0.000 | 0.484 ± 0.001 |
| | +M | 0.475 ± 0.004 | 0.478 ± 0.003 | 0.483 ± 0.000 | 0.480 ± 0.001 |
| | +CT | 0.484 ± 0.004 | **0.482 ± 0.002** | **0.468 ± 0.000** | **0.485 ± 0.000** |
| AUROC | C | 0.814 ± 0.001 | 0.781 ± 0.001 | 0.813 ± 0.000 | 0.802 ± 0.000 |
| | +ST | **0.826 ± 0.001** | 0.813 ± 0.000 | 0.824 ± 0.000 | **0.815 ± 0.000** |
| | +D | 0.824 ± 0.001 | 0.812 ± 0.001 | 0.824 ± 0.000 | 0.814 ± 0.000 |
| | +M | 0.821 ± 0.001 | 0.812 ± 0.001 | 0.823 ± 0.000 | 0.813 ± 0.000 |
| | +CT | 0.824 ± 0.001 | **0.815 ± 0.001** | **0.826 ± 0.000** | 0.815 ± 0.000 |
| NLL | C | 1.181 ± 0.004 | 1.505 ± 0.006 | 1.176 ± 0.000 | 1.216 ± 0.001 |
| | +ST | **1.148 ± 0.005** | 1.186 ± 0.001 | 1.147 ± 0.000 | 1.184 ± 0.001 |
| | +D | 1.158 ± 0.004 | 1.190 ± 0.005 | 1.149 ± 0.000 | 1.188 ± 0.001 |
| | +M | 1.164 ± 0.006 | 1.189 ± 0.002 | 1.151 ± 0.000 | 1.193 ± 0.001 |
| | +CT | 1.153 ± 0.006 | **1.173 ± 0.004** | **1.135 ± 0.000** | **1.180 ± 0.000** |
| ECE | C | **0.034 ± 0.003** | 0.087 ± 0.001 | 0.039 ± 0.000 | 0.055 ± 0.001 |
| | +ST | 0.037 ± 0.007 | **0.015 ± 0.003** | 0.034 ± 0.000 | 0.051 ± 0.002 |
| | +D | 0.044 ± 0.006 | 0.016 ± 0.002 | **0.034 ± 0.000** | **0.051 ± 0.001** |
| | +M | 0.047 ± 0.004 | 0.017 ± 0.006 | 0.039 ± 0.000 | 0.055 ± 0.001 |
| | +CT | 0.050 ± 0.005 | 0.019 ± 0.003 | 0.036 ± 0.001 | 0.057 ± 0.001 |

**Table 8 Performance across feature sets and models for Gender (2017/2019 train, 2021/2023 test split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.507 ± 0.003 | 0.500 ± 0.003 | 0.506 ± 0.000 | 0.508 ± 0.000 |
| | +ST | 0.508 ± 0.003 | **0.517 ± 0.002** | **0.524 ± 0.000** | 0.511 ± 0.001 |
| | +D | 0.511 ± 0.002 | 0.513 ± 0.002 | 0.523 ± 0.000 | 0.513 ± 0.000 |
| | +M | 0.507 ± 0.002 | 0.509 ± 0.003 | 0.523 ± 0.000 | 0.509 ± 0.001 |
| | +CT | **0.512 ± 0.006** | 0.515 ± 0.002 | 0.523 ± 0.000 | **0.514 ± 0.000** |
| AUROC | C | 0.551 ± 0.001 | 0.514 ± 0.001 | 0.546 ± 0.000 | 0.542 ± 0.000 |
| | +ST | **0.557 ± 0.003** | 0.544 ± 0.003 | 0.567 ± 0.000 | **0.547 ± 0.000** |
| | +D | 0.556 ± 0.003 | 0.540 ± 0.003 | 0.565 ± 0.000 | 0.540 ± 0.000 |
| | +M | 0.554 ± 0.003 | 0.540 ± 0.001 | 0.568 ± 0.000 | 0.540 ± 0.000 |
| | +CT | 0.556 ± 0.001 | **0.544 ± 0.002** | **0.570 ± 0.000** | 0.545 ± 0.000 |
| NLL | C | 0.961 ± 0.003 | 1.396 ± 0.008 | 0.976 ± 0.000 | **0.953 ± 0.001** |
| | +ST | 0.962 ± 0.003 | 1.002 ± 0.005 | 0.969 ± 0.000 | 0.954 ± 0.000 |
| | +D | 0.960 ± 0.006 | 1.002 ± 0.007 | 0.964 ± 0.000 | 0.954 ± 0.000 |
| | +M | **0.960 ± 0.004** | **1.001 ± 0.006** | **0.961 ± 0.000** | 0.954 ± 0.000 |
| | +CT | 0.963 ± 0.004 | 1.007 ± 0.011 | 0.962 ± 0.000 | 0.954 ± 0.001 |
| ECE | C | 0.056 ± 0.005 | 0.108 ± 0.003 | 0.058 ± 0.000 | 0.039 ± 0.002 |
| | +ST | 0.060 ± 0.004 | **0.057 ± 0.003** | 0.046 ± 0.000 | 0.041 ± 0.002 |
| | +D | **0.053 ± 0.003** | 0.059 ± 0.002 | 0.045 ± 0.000 | **0.034 ± 0.001** |
| | +M | 0.060 ± 0.006 | 0.062 ± 0.003 | 0.044 ± 0.000 | 0.039 ± 0.001 |
| | +CT | 0.060 ± 0.004 | 0.058 ± 0.002 | **0.044 ± 0.000** | 0.037 ± 0.002 |

**Table 9 Performance across feature sets and models for HH Income (2017/2019 train, 2021/2023 test split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.588 ± 0.001 | 0.542 ± 0.002 | **0.582 ± 0.000** | **0.586 ± 0.001** |
| | +ST | 0.588 ± 0.002 | **0.569 ± 0.002** | 0.579 ± 0.000 | 0.586 ± 0.001 |
| | +D | 0.587 ± 0.002 | 0.566 ± 0.002 | 0.579 ± 0.000 | 0.586 ± 0.001 |
| | +M | **0.588 ± 0.001** | 0.566 ± 0.001 | 0.581 ± 0.000 | 0.585 ± 0.001 |
| | +CT | 0.585 ± 0.002 | 0.562 ± 0.002 | 0.577 ± 0.000 | 0.581 ± 0.001 |
| AUROC | C | 0.606 ± 0.003 | 0.581 ± 0.000 | 0.605 ± 0.000 | 0.583 ± 0.000 |
| | +ST | 0.616 ± 0.004 | 0.610 ± 0.002 | 0.617 ± 0.000 | 0.599 ± 0.000 |
| | +D | 0.614 ± 0.002 | 0.610 ± 0.001 | **0.620 ± 0.000** | 0.598 ± 0.000 |
| | +M | 0.617 ± 0.002 | 0.610 ± 0.002 | 0.615 ± 0.000 | **0.604 ± 0.000** |
| | +CT | **0.618 ± 0.003** | **0.611 ± 0.002** | 0.613 ± 0.000 | 0.604 ± 0.000 |
| NLL | C | 1.219 ± 0.003 | 1.557 ± 0.022 | 1.228 ± 0.000 | 1.230 ± 0.000 |
| | +ST | **1.213 ± 0.005** | **1.244 ± 0.002** | 1.225 ± 0.000 | 1.224 ± 0.000 |
| | +D | 1.214 ± 0.005 | 1.244 ± 0.002 | **1.221 ± 0.000** | 1.225 ± 0.000 |
| | +M | 1.214 ± 0.003 | 1.244 ± 0.002 | 1.226 ± 0.000 | 1.222 ± 0.000 |
| | +CT | 1.219 ± 0.005 | 1.244 ± 0.002 | 1.225 ± 0.000 | **1.220 ± 0.000** |
| ECE | C | 0.063 ± 0.004 | 0.065 ± 0.004 | 0.067 ± 0.000 | 0.069 ± 0.001 |
| | +ST | 0.058 ± 0.002 | 0.080 ± 0.001 | 0.068 ± 0.000 | 0.070 ± 0.001 |
| | +D | 0.052 ± 0.008 | 0.079 ± 0.001 | 0.064 ± 0.000 | 0.072 ± 0.001 |
| | +M | 0.057 ± 0.006 | 0.080 ± 0.002 | 0.070 ± 0.000 | 0.072 ± 0.001 |
| | +CT | **0.047 ± 0.008** | **0.054 ± 0.002** | **0.043 ± 0.000** | **0.058 ± 0.001** |

**Table 10 Performance across feature sets and models for Number of Children (2017/2019 train, 2021/2023 test split). Values are mean±sd across folds; best per model in bold; best in metric in red.**

| Metric | Feature set | DNN | RF | GBM | SVM |
|---|---|---|---|---|---|
| Accuracy | C | 0.747 ± 0.002 | 0.741 ± 0.003 | 0.745 ± 0.000 | 0.754 ± 0.000 |
| | +ST | **0.756 ± 0.003** | **0.759 ± 0.001** | 0.749 ± 0.000 | 0.758 ± 0.001 |
| | +D | 0.751 ± 0.001 | 0.758 ± 0.002 | 0.747 ± 0.000 | **0.760 ± 0.000** |
| | +M | 0.752 ± 0.003 | 0.757 ± 0.001 | **0.750 ± 0.000** | 0.756 ± 0.001 |
| | +CT | 0.753 ± 0.002 | 0.758 ± 0.001 | 0.747 ± 0.000 | 0.757 ± 0.000 |
| AUROC | C | 0.819 ± 0.002 | 0.791 ± 0.001 | 0.819 ± 0.000 | 0.808 ± 0.000 |
| | +ST | 0.825 ± 0.002 | 0.809 ± 0.001 | 0.825 ± 0.000 | 0.814 ± 0.000 |
| | +D | 0.820 ± 0.001 | 0.811 ± 0.002 | 0.826 ± 0.000 | 0.814 ± 0.000 |
| | +M | 0.821 ± 0.001 | 0.810 ± 0.001 | 0.826 ± 0.000 | 0.814 ± 0.000 |
| | +CT | **0.826 ± 0.001** | **0.816 ± 0.001** | **0.828 ± 0.000** | **0.817 ± 0.000** |
| NLL | C | 0.704 ± 0.003 | 1.031 ± 0.016 | 0.712 ± 0.000 | 0.717 ± 0.000 |
| | +ST | **0.690 ± 0.002** | 0.729 ± 0.005 | 0.695 ± 0.000 | 0.705 ± 0.000 |
| | +D | 0.695 ± 0.002 | 0.722 ± 0.006 | 0.695 ± 0.000 | **0.704 ± 0.000** |
| | +M | 0.696 ± 0.005 | **0.720 ± 0.005** | 0.695 ± 0.000 | 0.706 ± 0.001 |
| | +CT | 0.696 ± 0.004 | 0.726 ± 0.004 | **0.692 ± 0.000** | 0.707 ± 0.001 |
| ECE | C | 0.025 ± 0.002 | **0.038 ± 0.004** | 0.020 ± 0.000 | 0.034 ± 0.001 |
| | +ST | **0.021 ± 0.004** | 0.040 ± 0.001 | 0.016 ± 0.000 | 0.023 ± 0.001 |
| | +D | 0.024 ± 0.003 | 0.040 ± 0.002 | 0.016 ± 0.000 | 0.022 ± 0.001 |
| | +M | 0.026 ± 0.003 | 0.041 ± 0.001 | **0.013 ± 0.000** | **0.017 ± 0.001** |
| | +CT | 0.026 ± 0.002 | 0.039 ± 0.001 | 0.015 ± 0.001 | 0.019 ± 0.001 |